STRESS MEASUREMENT AND REGULATION IN REAL-LIFE USING AFFECTIVE TECHNOLOGIES

by

Niaz Chalabianloo

B.S., Computer Engineering, Azad University, 2009

M.S., Computer Engineering, Middle East Technical University, 2013

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

Graduate Program in Computer Engineering

Boğaziçi University

2022

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Cem Ersoy, whose sincerity and encouragement I will never forget. With his immense knowledge and ample experience, he has been a role model, a true leader, and an inspiration to me throughout my PhD journey. I would like to extend my sincere thanks to the European Commission H2020 for supporting my work through Marie Sklodowska- Curie Innovative Training Network AffecTech: Personal Technologies for Affective Health. Additionally, thanks should also go to all the members of Netlab. Their kind help and support have made my study and life in Istanbul a wonderful time.

Furthermore, I would like to express my profound appreciation to my parents, Ghader and Malihe, my sister Farnaz, and my loving wife Camellia. Thank you for your constant love, support, and unwavering belief in me. I would also like to thank my friends Alper Alimoğlu, İhsan Mert Özçelik, Yekta Said Can, Deniz Ekiz, Ahmet Cihat Baktır, Can Tunca, and Muhammad Umair. Last but not least, I would like to thank all AffecTech early-stage researchers and supervisors for providing a fantastic support network and making beautiful memories.

ABSTRACT

STRESS MEASUREMENT AND REGULATION IN REAL-LIFE USING AFFECTIVE TECHNOLOGIES

Stress has become one of the main contributors to serious mental and physical health issues in today's world. Existing works in the literature have used Psychophysiological measures and proposed numerous mechanisms to detect stress and administer feedback to help users regulate it. Unobtrusive wearables' popularity is increasingly growing, intertwined with digital health notions, making them efficient, inexpensive, and easily accessible affective self-help technologies. This thesis first aims to investigate and implement stress detection mechanisms in the laboratory and everyday environments using unobtrusive wearable devices. In this regard, we investigate various scenarios, such as how to design and deploy stress measurement models that can efficiently use multi-modal data coming from different types of wearables used in the laboratory and real-life settings. We also study low-cost and practical methods for emotion regulation in stressful conditions of everyday life. In the next step, a mixed-methods study is conducted. For this, signals from multiple wearables and users' subjective opinions regarding different aspects of wearability were analyzed quantitatively and qualitatively. The next step is an in-depth study in cooperation with HCI researchers, in which we demonstrate the effects of haptic feedback on emotion regulation. As a next step for helping users choose the right device, we evaluate several wearables under completely identical conditions to compare the stress detection quality in wearables with different technologies. Finally, we utilize Explainable AI (XAI) to make our models more understandable for the end users, and in particular for the psychology and clinical experts. The results of our studies indicate that an integrated detection, notification, and intervention cycle is required to ensure a reliable system for regulating stress in daily life.

ÖZET

DUYGUSAL TEKNOLOJILERLE GERÇEK HAYATTA STRES ÖLÇÜMÜ VE REGÜLASYONU

Stres, günümüz dünyasında ciddi zihinsel ve fiziksel sağlık sorunlarına ana katkıda bulunanlardan biri haline gelmiştir. Literatürdeki mevcut çalışmalar, psikofizyolojik önlemleri kullanmış ve stresi tespit etmek ve kullanıcıların stresi düzenlemesine yardımcı olmak için geri bildirimi yönetmek için çok sayıda mekanizma önermiştir. Göze batmayan giyilebilir cihazların popülaritesi giderek artıyor, dijital sağlık kavramlarıyla iç içe geçiyor ve onları verimli, ucuz ve kolay erişilebilir etkili kendi kendine yardım teknolojileri yapıyor. Bu tez ilk olarak laboratuvar ve günlük ortamlarda stres algılama mekanizmalarını araştırmayı ve uygulamayı amaçlamaktadır. Bu bağlamda, laboratuvarda ve gerçek yaşam ortamlarında kullanılan farklı türde giyilebilir cihazlardan gelen çok modlu verileri verimli bir şekilde kullanabilen stres ölçüm modellerinin nasıl tasarlanacağı ve konuşlandırılacağı gibi çeşitli yolları araştırdık. Ayrıca günlük yaşamın stresli koşullarında duygu düzenleme için düşük maliyetli ve pratik yöntemler üzerinde çalışdık. Bir sonraki adımda, karma yöntemli bir çalışma yürütüldü. Bunun için birden fazla giyilebilir cihazdan gelen sinyaller ve kullanıcıların giyilebilirliğin farklı yönlerine ilişkin öznel görüşleri nicel ve nitel olarak analiz edildi. Bir sonraki adım, dokunsal geribildirimin duygu düzenleme üzerindeki etkilerini gösterdiğimiz derinlemesine bir çalışmadır. Kullanıcıların doğru cihazı seçmelerine yardımcı olmak için, farklı teknolojilerle giyilebilir cihazlarda stres algılama kalitesini karşılaştırmak için birkaç giyilebilir cihazı tamamen aynı koşullar altında değerlendirdik. Son olarak, modellerimizi son kullanıcılar ve özellikle psikologlar ve klinik uzmanlar için daha anlaşılır kılmak için Açıklanabilir Yapay Zeka'yı kullanıyoruz. Çalışmalarımızın sonuçları, günlük yaşamda stresi düzenlemek için güvenilir bir sistem sağlamak için entegre bir tespit, bildirim ve müdahale döngüsünün gerekli olduğunu göstermektedir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS/ABBREVIATIONS

ACC	Accelerometer
ANS	Autonomic Nervous System
BVP	Blood Volume Pulse
CBT	Cognitive Behavioral Therapy
CSV	Comma-Separated Values
CV	Cross-Validation
ECG	Electrocardiography
EDA	Electrodermal Activity
EDR	Electrodermal Response
EEG	ElectroEncephaloGraphy
EMA	Ecological Momentary Assessment
EMI	Ecological Momentary Intervention
EMG	Electromyography
ERM	Eccentric Rotating Mass
ERP	Event-Related Potential
ESR	Early Stage Researcher
FT	Fourier Transform
GBM	Gradient Boosting Machine
GSR	Galvanic Skin Response
HCI	Human-Computer Interaction
HRV	Heart Rate Variability
IBI	Interbeat interval
KNN	K-nearest neighbors
LDA	Linear Discriminant Analysis
LED	light-emitting diode
LightGBM	Light Gradient Boosting Machine
ML	Machine Learning
MLP	Multi layer Perceptron

Nasa-TLX	NASA Task Load Index
PNS	Parasympathetic Nervous System
PPG	Photoplethysmography
PSS	Perceived Stress Scale
RF	Random Forest
RFE	Recursive Feature Elimination
RFECV	Recursive Feature Elimination with Cross-Validation
SCL	Skin Conductance Level
SCR	Skin Conductance Responses
SCWT	Stroop Color and Word Test
sEMG	Surface Electromyography
SNS	Sympathetic Nervous System
ST	Skin Temperature
STAI	State-Trait Anxiety Inventory
SVM	Support Vector Machine
TSST	Trier Social Stress Test
XAI	Explainable AI
XGBoost	eXtreme Gradient Boosting
XTree	ExtraTree

1. INTRODUCTION

Failure to timely diagnose psychological and affective problems such as stress and being exposed to them for an extended period of time can lead to the emergence of more severe problems such as cardiovascular and physical health problems, depression, and becoming prone to other mental problems [1,2]. The long-term costs of these problems will be significant for both individuals and society, as well as healthcare systems and governments. Considering that millions of people suffer from these problems in today's advanced societies, it is easy to conclude that governments spend billions of dollars to deal with problems caused by stress. The most prevalent methods of managing stress revolve around finding ways to mitigate its adverse effects by identifying and alleviating it during and even before it takes place. In an efficient stress management system, the first and most crucial step is to detect the occurrence of stress and measure its fluctuations. Making any intervention, like informing the individual (awareness) to manage their stress (regulation) via specific instructions, comes next. A growing number of ubiquitous sensing devices have enabled monitoring of the vital body signals that allow human behavior, actions, and emotions to be predicted. It is now possible to record the physiological signals of the human body in order to analyze the biosignals representing the mental and emotional states with the help of ubiquitous mobile devices and wearables. It is noteworthy that most of these devices have relatively straightforward functionality and are easy to use for the end-user. There is, however, a significant challenge in identifying the meaning and concept of these signals and recognizing the types of affect they represent. There is a continuous effort among researchers to achieve the best possible results with more optimal algorithms and provide fully functional systems that can be presented to the end-user.

Affective computing emerged as an interdisciplinary field spanning cognitive science, psychology, and computer science. It involves the study of systems and devices that recognize, interpret, process, and simulate human emotion [3]. Computer science and engineering provide a full spectrum of tools and features to the area of "Affective

Computing". This includes designing and developing software and applications for subjective and objective data acquisition, data analysis, design and implementation of Ecological momentary assessment (EMA) and Ecological momentary intervention (EMI) tools, machine learning, and artificial intelligence, all essential for conducting evaluations and researches to improve affective health and mental well-being. Using the technologies offered by computer engineering, users can fill out EMA's and receive EMI's, receive biofeedback, practice emotion regulation and mindfulness, and most importantly, obtain the automatic diagnosis and prediction of their affective states, such as stress.

1.1. Research Questions

One of the main components of an affective health model can be outlined as a biofeedback mechanism that relies on the acquisition of physiological signals to evaluate a state of mind and convey that information to the user to facilitate its management. In biofeedback, biosignals are captured, and feedback is delivered using an output medium [4]. Through it, individuals can learn ways to adjust some of their body functions related to affective conditions such as stress to improve their affective well-being [5, 6]. A traditional approach to biofeedback relies mainly on audio and visual feedback in a controlled laboratory setting, where the user with sensors attached to their body is required to take a seat in front of a monitor screen [5]. However, as sensing technologies have matured and mobile devices have become more prevalent, biofeedback can now be provided using the actuators on mobile and wearable devices, such as visual, vibration, and even thermal actuations on unobtrusive wearable devices.

The key challenges this thesis attempts to address include how self-help technologies can help individuals capture, detect, comprehend and manage their emotions. We encountered many challenges in our quest to find reliable solutions to this problem. In order to find the right solution for these challenges, it was necessary to conduct repeated and additional testing and investigations both in the laboratory as well as in the real world. The questions and challenges are as follows:

- How to measure stress levels in real life with the help of wearable devices.
- How can contextual information be used for measuring the stress level in real-life using wearable devices?
- How to engage users in emotion regulation using smart stress detection mechanisms.
- Using a smart system, how can we overcome stress through affordable and convenient emotion regulation practices?
- How can we engage the end-users in the design and selection of sensors and actuators?
- How can we explain our stress detection model's decision-making process to the end user?

1.2. Thesis Outline

An introductory explanation of the causes and factors influencing stress and its adverse impacts on individuals' physical and mental well-being are presented in this first chapter, along with the thesis outline and its contribution to the body of knowledge.

In the second chapter, affects and emotions are defined. We explain the affect recognition using biosignals and describe the most effective biosignals used in our works and other types of signals utilized in similar works. Later, we briefly explain biofeedback and some of its common forms. Furthermore, the emotion regulation process and the impacts of utilizing methods such as yoga, mindfulness, and haptics for emotion regulation and stress reduction are also discussed in Chapter 2.

In the fourth chapter, we describe the standard methods and methodologies used in most of our research in detail. We discuss quantitative and qualitative data throughout this section, along with subjective and objective data types. In order to explain the subjective data, we examine the stress questionnaires used in our studies. In a comprehensive explanation of quantitative data, we describe how to collect information in the laboratory and daily-life environments and mention their advantages and

disadvantages. We also examine the differences between physical and mental stress in laboratory settings and explain several methods of stress induction. Moreover, we describe the third-party programs we use to analyze the raw signals and conclude the methodology chapter by describing the types of sensors we use.

In Chapter 5, we present an unobtrusive smart stress detection mechanism suitable for daily life, capable of suggesting appropriate relaxation methods such as yoga practices or mobile relaxation applications for emotion regulation.

In the sixth chapter, we explain our experiment on the potentials of using vibration and heat haptics for emotion regulation. We also explain how users were engaged in personalizing their haptic patterns. The chapter concludes with a statistical analysis of the quantitative results in terms of the feeling perceived by the users, as well as the differences between vibration and thermal haptics.

In Chapter 7, several different wearables are analyzed in detail in a mixed-methods approach, both qualitatively and quantitatively. In Chapter 8, more wearables are examined in stress detection and measurement application. In Chapters 7 and 8, we describe how to minimize the effect of environmental noise on PPG sensors in this chapter after reviewing various data preprocessing techniques. Next, we discuss how to prevent conditions that can lead to unrealistic and biased results and data leakage. Additionally, we examine how multimodality affects model accuracy in a significant manner. To increase the human-centric nature of our study, we use SHAP to interpret our machine learning models and explain how our final models become understandable for end users, clinicians, and psychologists.

2. BACKGROUND

2.1. Detecting Affect using Biosignals

An overview of sensing technologies for detecting affect is provided in this section. The works presented in this dissertation have exclusively employed wearable sensors, mostly with skin conductance, such as Electrodermal activity (EDA) and heart rate variability (HRV) sensors, to investigate the development and regulation of emotions. All of the employed wearables equipped with these two types of sensors must be in close contact with the skin in order to properly capture biosignals from key body locations. Due to the fact that these wearable and unobtrusive devices do not entail multiple electrodes, they are almost effortless and straightforward to wear and take off and can be utilized in a daily context.

2.1.1. Autonomic Nervous System (ANS)

The autonomic nervous system carries out control of the body's unconscious actions. The autonomic nervous system (ANS) is a branch of the peripheral nervous system that influences the activities of the body's internal organs [7]. A large number of involuntary bodily functions are controlled by the autonomic nervous system, including heart rate, blood pressure, respiration, and pupillary response. It consists of three anatomically distinct components: sympathetic, parasympathetic, and enteric nervous systems. However, depending on the source, the last one may be considered a part of the autonomic nervous system or an independent system. It is often believed that the sympathetic nervous system (SNS) is responsible for the hormonal and neuronal stress response, commonly referred to as the "fight or flight". In contrast, the parasympathetic nervous system (PNS) is responsible for "rest and digest". These two systems often work in opposite directions where one activates physiological responses, and the other inhibits them (see Figure 2.1).

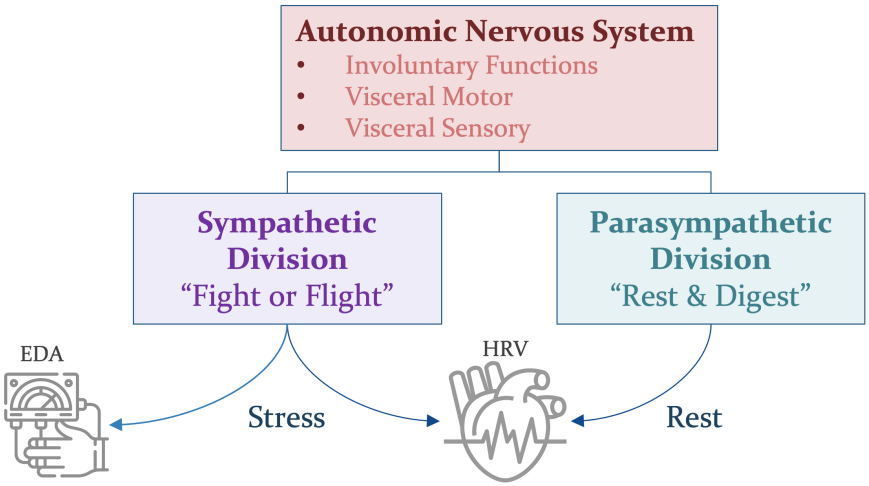


Figure 2.1. Sympathetic and parasympathetic components of the ANS.

2.1.2. Cardiovascular Activity

For electrocardiography (ECG) or photoplethysmography (PPG) sensors to work, it has to sense electrical impulses generated by the heart's beating and blood flow within the body. The process of recording the heart's electrical activity is referred to as an electrocardiogram (ECG). Electrograms indicate the heart's electrical activity by graphing the voltage versus time using electrodes attached to the skin surface. These electrodes detect small electrical changes caused by cardiac muscle depolarization and repolarization during each heartbeat [8, 9].

The term "ECG" is traditionally used to refer to a 12-lead ECG taken while in a supine position. However, there are also other devices that can record ECG, such as Holter monitors and even smartwatches capable of recording ECG. Signals from ECGs can also be recorded with other devices in other contexts. An ECG has three main components: the P and T waves, and the QRS complex, representing the depolarization of the atria, repolarization of the ventricles, and depolarization of the ventricles, respectively. Heart activity is a prominent signal for distinguishing affect due to the direct influence the autonomous nervous system (ANS) has on heart rate.

2.1.2.1. Heart Rate Variability (HRV). When the heart beats, it triggers an electrical impulse which can be captured by biosensors. In order to capture the HRV, PPG and ECG biosensors incorporate technologies that are totally different. A heartbeat is measured by an ECG sensor using electrodes on the body, whereas blood flow is measured by a PPG sensor using light-based technology [10]. These sensors measure heart rate in beats per minute (BPM), which are not always at a constant frequency.

The measure of the variation in the time interval between two consecutive heartbeats is called heart rate variability (HRV) [11]. Inter-beat interval (IBI), NN interval, and peak-to-peak interval, with the last belonging to PPG and the first three belonging to ECG, are also terms used to describe these variations, which all are measured in milliseconds. HRV can be measured in the frequency, time, and nonlinear domains [10], [12]. The most common frequency-domain features consist of HF (high-frequency component), LF (low-frequency component), and LF/HF (the ratio of LF to HF). Time-domain features consist of STD RR (standard deviation of the inter-beat interval), Mean RR (mean value of the inter-beat intervals), NN50 (the number of pairs of consecutive NNs that differ by more than 50 ms), pNN50 (percentage of consecutive beat-to-beat intervals that vary by more than 50 ms), and, RMSSD (root mean square of successive differences of the R-R intervals). Nonlinear features include sample and approximate entropy and multiple components of Poincaré plots. HRV is reported to be an indicator of the activity of both the parasympathetic and sympathetic nervous systems. Its measurement can be done from a single sensor, allowing it to be used in daily life to regulate autonomic balance [11], [13]. In general, a high HRV reflects effective emotion regulation. In contrast, a low HRV reflects states of stress and anxiety. However, these assumptions are different for distinct components and features of HRV. For instance, high anxiety, stress, and excessive time pressure have been shown to decrease high-frequency (HF) activity [14, 15].

2.1.2.2. Blood Volume Pulse (BVP). Using a photoplethysmography (PPG) sensor, BVP measures the heart rate based on the blood volume that passes through the tissues in a localized site with each heartbeat. There is a potential measurement site

wherever a pulse can be easily accessed, but fingertips and earlobes are more commonly used. Biofeedback training often uses BVP rather than ECG. The latter is likely more preferable in some clinical situations or situations where the subject is prone to make lots of movements. There are some potential measurement errors with the BVP sensor, which makes it less precise than the ECG but easier to apply for biofeedback training applications.

The PPG sensor shines infrared light onto the body surface, primarily by a light-emitting diode (LED). A PPG sensor transmits this light through the tissues that backscatter and reflect it to the PPG sensor's photodetector [16]. Hemoglobin in the red blood cells selectively absorbs red light while other tissues reflect it, which explains why the technology works. As the relative blood volume in the tissue increases, so does the amount of light returning to the PPG photodetector. Blood flow is represented by the BVP amplitude, which is derived from the raw BVP signal. Despite the fact that BVP features captured by PPG sensors can be used independently, they are generally used to derive heart rate variability (HRV) features that can be utilized to detect stress levels [17–19].

2.1.3. Electrodermal Activity (EDA)

Although the idea of using Electrodermal Activity (EDA) in psychological research has been around for a long time [20], it is still among the top biosensing measures employed in the subject area of affective computing using ubiquitous and wearable devices [21, 22].

Electrodermal activity (EDA) is an umbrella term for Galvanic skin response (GSR), which measures the skin's electrical properties related to the autonomic nervous system's activation. Electrodermal activity is a phenomenon when physiologically provoking events result in improved electrical conductivity of the human skin. An EDA sensor measures the electrical conductivity of the skin, which is a result of SNS activity that can be triggered by either internal or external emotional stimuli [23].

Skin resistance is believed to vary with the state of sweat glands in the skin according to the traditional theory of EDA. Psychological or physiological arousal influences sweating through the sympathetic nervous system [24]. In response to highly aroused sympathetic nerves, sweat gland activity will increase, thereby increasing skin conductance. Hence, skin conductance can be used to measure emotional and sympathetic responses [25]. Tonic sympathetic activity and fast phasic sympathetic activity are both evident in the EDA. The electrodermal level (SCL) is a unit of tonic activity, while the electrodermal response (EDR) is a unit of phasic activity [26]. Phasic parameters determine tonic changes (EDL). Tonic EDA can be evaluated based on spontaneous fluctuations of nonspecific EDR. In particular, the frequency of nonspecific EDR during a particular time period can be used as an indicator of EDA. In studies of general alertness and arousal, tonic EDA is found to be beneficial [26]. As part of the EDA (Phasic), Skin Conductance Responses (SCR) represent the faster and event-related components. Tonic (SCL) is used for calculating the baseline and extracting statistical features such as standard deviation, percentile, and mean, as it does not contain peaks that will affect the calculation of the baseline.

It has been found that EDA, along with the heart rate signal, is one of the best discriminating signals in affective computing and emotion detection research. Various methods in the literature have been used to measure stress in an individual using EDA. The most common EDA features utilized in studies such as stress detection include standard deviation, minimum and maximum values, mean amplitude, the delay between the application of stimuli to the user and the response, number of peaks and their height, and rising and recovery times [27]. Some examples of biosensor devices that are primarily wearables and have been used in the literature for EDA measurement are Empatica E4, Microsoft Band 2, BITalino biosensing platform, and Shimmer3 GSR+ [28, 29].

2.1.4. Respiratory System Activity

Sensors that monitor respiration (or breathing) are used to measure the number of inhalations and exhalations during a breathing cycle, which facilitates the gas exchange in the lungs during the process of breathing. As soon as the lungs are filled with air, the air is forced out of the lungs at the end of every breathing cycle. Breathing signals reveal the dynamics of respiration, that is, the process regulating gas exchanges in the lungs and supporting speech and sound. Through monitoring this fundamental function of breathing, we can achieve insight into problems related to apnea, oxygen intake, metabolic changes associated with physical activity, and breathing responses to psychological stress. There is a close connection between respiration and the cardiovascular system. Anxiety, for instance, can cause a shallower and faster respiration rate, and breathing rates can be influenced by stress and excitement [30]. Also, there is a close relationship between respiration and HRV, which is one of the essential biosignals in identifying stress and emotions. The results of existing studies show that slow breathing, which occurs at a rate of 5.5-8 breaths per minute, is related to a higher heart rate variability (HRV), which is an essential indicator of calmness [31]. Respiration sensors record the inhalation and exhalation cycles of breathing. A respiration sensor attached to a subject's body is referred to as a contact-based measurement. A piezoelectric abdominal band is one of the most commonly used contact-based methods. During abdominal breathing, the sensor produces a signal as a result of stretching an elastic material.

Additionally, other contact-based methods can be used to collect different measurements such as humidity and temperature of the air, respiratory airflow, and sounds to assess an individual's inhalation and exhalation cycles [32]. Unlike the contact method, the non-contact method measures chest displacement using an infrared or proximity sensor. It is crucial to keep the subject still during data collection since both types of sensors are prone to noises caused by body movement, coughing, and talking. Breathing signals are analyzed for features such as respiration rate, breathing amplitude, and duration of inspiration and expiration.

2.1.5. Other Physiological Signals Used for Affect Recognition

While almost all of the aforementioned physiological signals or a combination of some are utilized in our studies, there are also other physiological signals, such as Electroencephalography (EEG) and Electromyography (EMG), that are used in emotion recognition studies. However, we did not include these signals in our studies simply because the commercially available wearables equipped with these sensors are not sufficiently unobtrusive and cannot be easily used in daily life, while one of our research's main objectives is to provide mechanisms that can be used in daily life and on a regular basis.

2.1.5.1. EEG (ElectroEncephaloGraphy). Electrical activity in the brain is measured by electroencephalography (EEG). Using EEG, a complex overview of neural activity oscillations is obtained by non-invasively collecting signals from various standard scalp locations. Single and multiple channel electrodes are used in EEG bands to measure electrical signals corresponding to neural activity. For the purpose of maintaining the electrical connection, the electrodes of an EEG sensor require direct and close contact with the skin, which is achieved using headbands or adhesive gel. An EEG band records changes in neural activity in response to stimuli. Two stages of signal analysis are performed regularly on EEG signals: preprocessing and postprocessing. Raw signals are cleaned through preprocessing, which removes artifacts by filtering the data. As a next step Fourier Transform (FT) can be used to extract features that machine learning algorithms will utilize for classification [33]. In addition to being complex and generally fast, EEG signals are also prone to head and eye movements.

A reduced channel wearable headset is usually lightweight and easy to wear and take off of the head, making it less obtrusive. However, it lacks the signal resolution and precision of a traditional multichannel headset. Behavioral or psychological states are reflected in the EEG using delta, theta, alpha, which indicates a balanced and calm state of mind and a reduction in stress, and beta waves, associated with cognitive and emotional processes and an increase in response to stress. There are different frequency

bands for each of these waves (0.1 to 100 Hz). Stress can be detected by analyzing EEG mean amplitude, mean of Event-Related Potential (ERP) amplitudes, and theta, beta, and alpha frequency bands.

2.1.5.2. sEMG (Surface Electromyography). The contraction of human muscles is caused by electrical stimulation by neural signals. Surface Electromyography (sEMG) refers to recording the electrical activity produced by skeletal muscles. By using surface electrodes that capture the potentials of the fibers they lay upon, the electrical activity of the muscles (voltage over time) can be easily recorded, a procedure that was traditionally performed by an invasive needle electrode (intramuscular EMG). An electromyographic signal (sEMG) is produced by this measurement, which provides information on the motion and biomechanics of the contracted muscles. Signals generated by electromyography provide information about the contraction of specific muscles of the body. Signals generated by electromyography represent rapid voltage oscillations in time, with the approximate amplitude range of 5 mV.

As a result of a common EMG signal analysis, different aspects can be assessed, such as the duration of muscle contractions, the specific timing in which movements or contractions are occurring, and fatigue or muscle tensions. In examining affect, startle reflexes have been used as an example of the body's reaction to strong and intense stimuli. The trapezius and facial muscles are particularly relevant when measuring muscle activity for emotion recognition. An EMG electrode placed on the face can be used to measure the human body's responses to an unexpected and strong stimulus in the form of actions such as quickly blinking and contracting different muscles across the body. For instance, Sioni et al. utilized an electromyogram (EMG) sensor to detect stress by measuring facial muscle activity [34]. Firm contact is essential for accurate EMG measurements. Motion artifacts can be introduced to the EMG signal when muscle movement is present, and activities such as talking or coughing can impair the assessment of muscle activity.

2.2. Detecting Affect Using Non-biological Signals

2.2.1. Accelerometer

Researchers have demonstrated that different emotional states can be detected by observing movements of the human body and postures. Castellano et al. used multimodal data to examine the dynamics of body movement in order to identify human affective behaviors. Several movement metrics were used to determine emotions, and the amounts of movement, intensity, and fluidity were shown to be key factors in determining the type of emotion [35]. Melzer et al. examined whether movements comprised of collections of Laban movement components could be interpreted as representing basic emotions [36]. Their study confirms that, even when the subject has no intention of expressing their emotions, specific movements can aid in the perception of bodily expressions of emotions. Therefore, movements and affects may be detected by accelerometer sensors.

2.3. Feedback to Affect

Mechanisms for generating actuation as a form of feedback to humans play an essential role in creating a complete feedback interaction mechanism. This includes sensing the physiological and psychological states of the individuals' bodies and making them aware. There are several different types of actuators, but in general, they are comprised of mechanical components that are mechanically driven in response to input signals for the purpose of controlling a system or providing information about it. We believe that it is essential to emphasize that when it comes to studies related to affect recognition and emotion regulation, the focus must be on actuation technologies that can be conveniently implemented and coupled with the human body. A variety of modalities can be used to provide feedback on affective states [37–39]. Nowadays, biofeedback is provided to users through visual and shape-changing feedback or sound, temperature, and vibrotactile haptics [29]. In affective computing, a similar approach can be utilized to explain biosignals and their indications to the user and employ

effective mechanisms to control actuators to provide biofeedback to the users. In a fully functional architecture for identifying emotion and performing its regulation, a complex set of mechanisms must work together to decide on how to interpret the biosignals and when, how, and where this feedback must be issued. Since addressing all actuators in detail is beyond the scope of this thesis, we will only briefly explain those we have already employed in parts of our research. It must be noted that in a part of our studies where we did close collaborations with researchers from the Human-Computer Interaction (HCI) community, only vibration and temperature actuators were used. Therefore, in this part, we will introduce only these two cases as well as display actuators that are the most widely used actuators in biofeedback.

2.3.1. Visual Biofeedback

The purpose of a biofeedback system based on a screen is to provide information regarding changes in the body that have occurred over time. Its purpose is to provide the researcher with a way of assessing the dynamics of the changes mentioned above, therefore providing a means of gaining a better understanding of and tracking the inner state of a particular subject. Typically, these types of techniques are used to track health metrics or sports performance. Examples include ECG feedback, respiration feedback, and movement tracking. Biosensors based on screens are commonly used in clinical settings and in hospitals as a means of providing feedback. The psychology field has adopted biofeedback as a technique to self-regulate emotions, as research shows that the technique contributes to the improvement of emotional self-regulation.

Traditional visual biofeedbacks that are based on screen commonly utilize two-dimensional graphics and incorporate elements such as colors, patterns, and lights to display a signal that is constantly changing over time [40, 41]. Whenever the representation is updated along the time axis, the peak and trough of the signal appear in an axis showing the measurement magnitude in a particular range so that both sharp and gradual dynamics can be observed. Currently, the research direction has begun exploring alternative visual technologies that are not based on traditional screens and displays

by utilizing materials' experiential qualities and aesthetics [42]. Visual displays that do not rely on screens include ambient light, electroluminescent, and thermochromic displays [43]. Unlike screen-based technologies, both are flexible and thin and can be fabricated in a wide range of shapes through multilayer fabrication [44].

2.3.2. Haptic and Temperature Actuation

The term haptic technology refers to any technology which can give the user a tactile experience using vibrations or motions. Electronics that provide haptic feedback typically utilize vibrations, and most employ an eccentric rotating mass (ERM) actuator comprised of an unbalanced weight connected to a motor shaft. This irregular mass spins as the shaft spins, causing the actuator and the coupled component to vibrate [39]. Such tactile vibrations coupled to a biosensor can be used to provide information about a biosignal being tracked. Researchers have also used temperature (Heat/Cool) to give haptic feedback alongside vibrotactile feedback. For instance, when a current is applied to a heat-resistive material, it produces heat, which can be used as temperature feedback [45].

2.4. Emotion Regulation

2.4.1. Affects and Emotions

The perception, interpretation, and interaction with the environment around us are profoundly influenced by a vital part of our everyday lives i.e. affect [46]. Affect is the psychological term used to describe the underlying perception of moods, feelings, and emotions. The process of effectively managing your emotional responses is referred to as affect (emotion) regulation [47]. An individual who is able to regulate their high arousal negative affect, for example, by reducing their arousal, will be able to improve their overall psychological well-being and affective health. Similarly, the inability to moderate an individual's emotional responses can lead to its deterioration [48].

Managing one's emotions is a multifaceted procedure that relies on initiating, inhibiting, or modulating individuals' state of mind in a particular condition. Emotion regulation aims to increase pleasant emotions (joy and happiness) while decreasing unpleasant emotions (sadness, fear, anger). Additionally, emotion regulation refers to mechanisms such as the ability to concentrate on a task and the capability to restrain inappropriate behavior.

2.4.2. Emotion Regulation Process Model

Emotion regulation occurs in five phases (see Figure 2.2) in the following order:

- Situation selection
- Situation modification
- Attentional deployment
- Cognitive change
- Response modulation

The "situation selection" refers to making a decision regarding whether to avoid or approach a situation that is emotionally relevant. Whenever a person avoids or disengages from such a situation, they decrease the likelihood that they will experience an emotion [47]. By contrast, individuals who choose to approach or engage with an emotionally relevant situation increase their chances of experiencing an emotion. The concept of "situation modification" is the process of modifying a situation to change

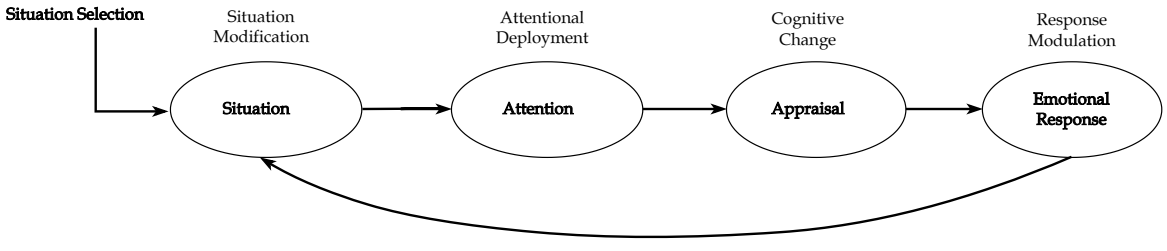


Figure 2.2. James Gross's emotion regulation model for stress management.

that situation's emotional impact, for instance, adding humor to a conversation in order to induce laughter. This term specifically refers to the alteration of the individual's external circumstances, while alteration of an individual's internal environment in order to change the person's perception of a situation to alter its impact is referred to as "cognitive change" [49]. An attentive deployment involves directing the person's attention toward or away from an emotional event. Last but not least, "Response modulation" is concerned with attempting to influence the behavioral, physiological, and experiential elements of the affective response systems directly [47].

2.4.3. Emotion Regulation and Stress

Whenever faced with stressful events, people respond by making autonomic and coordinated efforts to minimize the negative consequences and maximize the positive effects. It can influence what emotions they have when they have them and how they experience and express those emotions. An individual's ability to influence what emotions they feel, when they feel, and how they experience and express them is known as emotion regulation, which may be defined as the act of regulating emotions [47]. Various studies have suggested that the concept of emotion regulation is a broad term that encompasses the regulation of all emotional reactions that are triggered, from the simplest emotions to various mood states, in addition to regulating daily living [47]. It is possible to minimize stress through a variety of interventions, depending on the individuals' preferences. A number of stress alleviation practices have been cited as being helpful in combating stress, including ancient practices such as yoga [50], and other physical activities. In the same way, meditation, meditative awareness (mindfulness) [51], breathing exercises, relaxation techniques, and cognitive behavioral therapy (CBT) [52] are all proven to be beneficial [53, 54]. However, some of these techniques cannot be used in offices, social settings, or most daily activities. Thus, a stress management application based on a smart device may be of great use.

In recent years, several smartphone applications have been developed for this purpose, including but not limited to Calm, PAUSE, Heartmath, and Sway. However,

these applications do not include biofeedback and are not customizable for personalization, and only a limited number of studies have validated their effectiveness [55].

Many psychological scientists have studied perceived stress. When contextual demands and perceived resources fail to match constantly (as opposed to occurring at a specific moment), an individual is said to be experiencing chronic stress. The effects of chronic stress can be seen not only in people's well-being and quality of life but also in the onset and progression of several physical and mental diseases [56]. Consequently, researchers have been exploring how people alleviate perceived stress's physical and cognitive burden through different mechanisms. Coping mechanisms, stress management strategies, self-regulation, or emotion regulation practices are different terms that describe the methods by which people implement specific behavioral, cognitive, or emotional methods for allosteric regulation [57].

Individuals' ability to regulate emotion is directly related to their response to stress. In spite of the fact that stress management and emotion regulation differ in numerous ways, both concepts require modulation of affect and appraisal of emotion [58]. In a 2022 study, Griffin et al. showed that physiological responses to stress in laboratory settings have also been associated with emotion regulation [59]. Self-regulating activities can also be used to reduce psychological stress in various situations, such as academic examinations [60]. For instance, when self-regulatory actions are employed before a school exam, levels of emotional strain are significantly reduced, which may lead to better academic results.

2.4.3.1. Emotion Regulation Through Yoga and Mindfulness. Yoga was developed in ancient India more than 2000 years ago. It is a discipline or group of physical and mental practices aiming to manage, calm, and regulate emotions and the mind. Over time, yoga evolved to incorporate physical movements in the form of postures, which were integrated into its traditional breathing and relaxation practices. Yoga's primary goal is to improve human well-being by creating physical flexibility and alleviating pain and unpleasant thoughts and feelings. Both mental and physical health conditions,

such as anxiety, depression, and cardiovascular disease, have been reported to benefit from yoga. It is widely practiced in many different forms around the world and has become a global trend. Relaxation is an integral part of all types of yoga. Furthermore, some forms focus primarily on pranayama (focusing on the breath), while others are more physical. Yoga practices such as vinyasa (individual poses linked by flowing movements) involve using the breathing pattern to move through various postures. These movements become meditative when done correctly. The practice commonly includes pranayama, standing postures, and vinyasa. In addition to increasing fitness and flexibility and maintaining their linkage to breath, vinyasa helps to keep the body moving. Besides seated postures, the practice may also include inversions, and a final relaxation referred to as savasana [61].

Being mindful involves paying attention to the here and now rather than focusing on the past or future. It is often referred to as being present. In addition to being aware of what we consume as food, as well as physical stimuli such as feeling the wind on our hair, being present can include paying attention to our surrounding environment. As part of mindfulness, we acknowledge our thoughts and bodies. Thousands of thoughts pass through the minds of humans each day, many of them without any consequence. Sometimes, these thoughts are repetitive and negative in nature, leading to increased stress and unpleasant physical symptoms like anxiety. In order to be mindful, we must be aware of our thoughts and whether we are caught up in them rather than being present in the moment. Additionally, being mindful involves becoming more connected to the sensations in the body by becoming more aware of the physical body on a daily basis. This experience may include sensing the legs swaying while walking or sensing the ground beneath the feet. Mental and physical health have both been demonstrated to benefit from mindfulness. The National Institute for Clinical Excellence recommends it as an adjunct therapy to Cognitive Behavioural Therapy (CBT) for preventing depression relapses [62]. However, there are various distractions around us that may make this difficult for some individuals. In such a case, individuals can choose a convenient time and location to start becoming aware of their breathing and body sensations as they sit in a comfortable position.

2.4.3.2. Mobile Applications for Emotion Regulation. Excessive use of smartphones can harm individuals' mental health [63], which is utterly contrary to our goal of emotion regulation and developing mindfulness. The prevalence of smartphones in the modern world has raised questions about the feasibility of their use in regulating emotions and practicing mindfulness. Therefore, rather than being pessimistic about smartphone use and considering it always harmful, we should consider how we can employ our smartphones to benefit from mindfulness exercises. While it is impossible to practice physical emotion regulation practices like yoga anywhere and anytime, individuals can reduce stress and anxiety by utilizing their mobile phones to perform mindfulness exercises. Several mindfulness apps are available for smartphones that can be used to guide people in their daily mindfulness practice. One of them used in a part of studies related to this thesis is PAUSE. Utilizing the PAUSE application, users can practice focused attention while on their mobile phones. Mindfulness and Tai Chi principles are the basis for how PAUSE works, and its users can easily start practicing relaxation whenever and wherever they want. While using PAUSE, in addition to receiving calming audiovisual feedback from it, users must also move their fingertips slowly and continuously across the mobile phone screen. These practices help the body's parasympathetic nervous system to stimulate rest and digest responses and help the users quickly reduce their stress levels.

2.4.3.3. Haptics for Emotion Regulation. The use of haptics to modulate response includes vibrations imitating a slower heartbeat or vibrations generated at a rate of 60 bpm to assist users in regulating their response during stressful situations [64, 65]. Researchers have found that vibration administered 30% below baseline heart rate can reduce an individual's anxiety and stress and elevate their HRV under a stressful situation. In contrast, an increase in HRV and subjective anxiety was observed when fast feedback was offered at frequencies 30% greater than the baseline heart rate. [66]. As well as vibrations mimicking the heart rate, there are existing researches that use vibrations to regulate emotions by means of slow breathing, which helps improve heart rate variability, which is a key indicator of adaptability [67, 68].

This growing interest in haptics in the human-computer interaction community has primarily focused on vibrotactile actuators as a means of helping users with emotion regulation [64, 65], [69]. Nonetheless, emotion regulation with thermal feedback has been less studied regarding its material and experiential aspects. In a study by Jonsson et al., heat is investigated experientially as a material for design, and its use has been explored through a user study [70]. They reported that thermal cues have a subjective nature, as different people report different levels of sensitivity and appreciation of it with totally different acceptable ranges. Additionally, the authors indicate that heat, as opposed to other haptic modalities such as tactile or vibration modalities, can be perceived deep within the body and as comfortable and subtle [70].

3. RELATED WORK

3.1. HRV and Biofeedback

There is an increasing interest in research on affective health and well-being, and affective technologies are being developed to treat stress-related disorders in adults, youth, caregivers, healthcare workers, and students [29], [71]. These technologies aim to prevent, diagnose, triage, intervene, self-manage, and maintain affective disorders in clinical and non-clinical settings. HRV biofeedback is one of the essential functions of interactive systems for affective well-being (see Figure 3.1).

The HRV is influenced by activities such as physical exercise, eating, and sleeping. Additionally, it is strongly linked to emotional arousal and decreases during emotional stress. In particular, this is of significance because parasympathetic activity, also known as vagal tone, is involved in processes of self-regulation needed for psychological, emotional, and affective well-being [72]. It is possible for both High and Low frequency (HF, LF) features of HRV to be influenced by various circumstances. HF is thought to be affected by parasympathetic activity based on existing research on

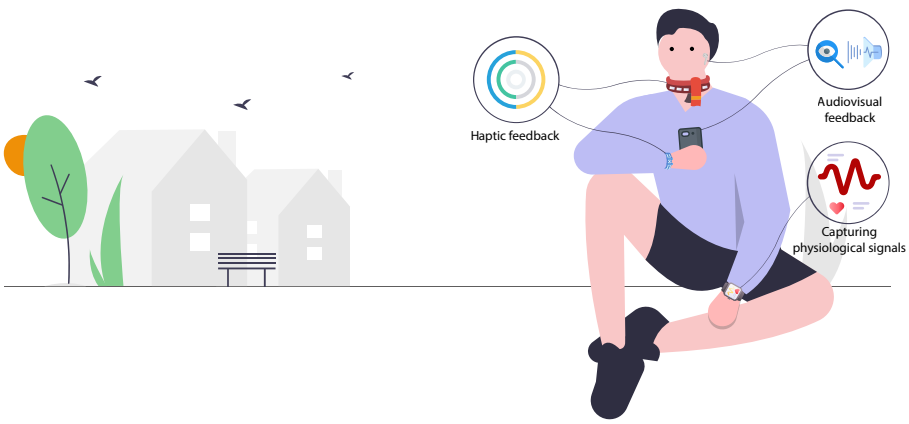


Figure 3.1. Monitoring biosignals and providing biofeedback via visual, auditory or haptic mechanisms.

Table 3.1. Heart rate variability features and their definitions.

Feature	**Description**
Mean RR	Mean value of the inter-beat (R-R) intervals
STD RR	Standard deviation of the inter-beat interval
RMSSD	Root mean square of successive differences of the R-R intervals
pNN50	Percentage of the number of successive R-R intervals varying more than 50ms from the previous interval
SDSD	Related standard deviation of successive R-R interval differences
TINN	Triangular interpolation of R-R 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)

HRV components [73, 74]. At the same time, LF is considered to be indicative of both sympathetic and vagal activity [10]. Boonnithi et al. investigated the HRV features and found the LF, mean RR, and the difference between LF and HF to be the most distinctive features for detecting stress [75]. Time-domain features such as RMSSD and pNNx are more closely correlated with parasympathetic activity compared to the standard deviation of NN intervals (SDNN) [12]. Table 3.1 lists some of the HRV features most commonly used in stress detection and biofeedback studies. Through biofeedback, HRV can not only be utilized to measure regulatory efficiency but also used to restore regulatory flexibility [11]. A biofeedback system uses biosensors to monitor internal bodily processes and provide feedback to individuals so they can actively control their physiological functions [76]. Few exemplary systems include those for self-awareness and reflection [77, 78], emotion regulation [64, 79], as well as those focused on the role of the body for relaxation or mindfulness exercises [80, 81]. As a result of its effectiveness, HRV biofeedback has been employed in various applications [82], [13]. Research on stress biofeedback has shown that SDNN decreases during stress and RMSSD increases during biofeedback-assisted breathing [43].

There is a wide range of modalities for providing HRV biofeedback. In most cases, feedback is given through visual [79], auditory [83], or haptic [77], [64] changes, which have been shown to regulate bodily functions and improve well-being in stress-related and affective situations [84, 85]. As illustrated in Figure 3.1, HRV biofeedback can be delivered visually either via a smartwatch or a mobile phone screen, while the haptic feedback can be provided through a vibrotactile actuator present in some devices. An external actuator can be used to deliver HRV biofeedback if the feedback modality is not present in the device [77]. In order to provide reliable HRV biofeedback, the sensor data must be accurate, and the device should be appealing to the user in terms of unobtrusiveness, wearability, and comfort.

3.2. Emotion Regulation

In a study by Ahani et al., the physiological impacts of mindfulness were investigated. Their experiments were conducted using the Biosemi device that acquires electroencephalograms (EEGs) and respiration signals. With machine learning algorithms, they were able to distinguish control states (non-meditative) from meditation states [86]. A wearable EEG measurement device (Muse headband) was used by Karydis et al. to identify the post-meditation perception states [87]. Mason et al. studied how yoga affects physiological signals [88]. They measured respiration signals and blood pressure using a PortaPres Digital Plethtsmograph. Through the use of these signals, they were also able to demonstrate the positive effects of yoga. Another study validated yoga's positive effects with physiological signals; researchers used a piezoelectric belt and a pulse sensor to monitor breathing, and heart rate signals [54]. Their study showed the beneficial impact of different yoga breathing patterns on helping subjects relax. Using physiological signals in mobile mindfulness apps has also been shown to be effective in several studies. A few variables were monitored by Svetlov et al., including heart rate variability (HRV), electrocardiographic activity (EDA), salivary alpha-amylase (sAA), and electroencephalography (EEG) [53]. The effect of mobile mindfulness apps has also been validated using EEG and respiration signals [89]. Upon reviewing the literature, it is possible to observe that traditional relaxation practices and mobile mindfulness

methods are investigated individually in separate studies. Since people spend more time in office-like environments in the modern age, traditional methods are not suitable for most people due to their out-of-office environment requirements. In contrast, some smartphone applications, such as PAUSE, HeartMath, and Calm, can be used in an office setting. These apps do not require any additional equipment or hardware, which makes them ideal choices for indoor settings.

3.3. Comparison and Validation of HRV Monitoring Devices

The purpose of this section is to summarize the quantitative and qualitative studies that have been conducted for the purpose of comparing HRV data quality and usability of wearable devices employed to measure HRV.

3.3.1. Quantitative Comparison Studies

Medical-grade types of equipment were used as the reference golden standard in the first experimental practices conducted in laboratory settings. These devices delivered ground truth data and functioned as the foundation for building blocks of comparative studies for assessing the usability, reliability, and validity of wearable biosignal tracking devices.

There have been many experimental and survey studies investigating the accuracy of photoplethysmography (PPG) as an estimate of HRV, and how it could be used as a surrogate of electrocardiography (ECG) [90–93]. In one of the earliest comparative studies conducted in a 2006 study performed for the validation of PPG and ECG readings, Yu et al. [92] proposed a system for automatically identifying the reliability of heart rate measurements using a combination of PPG and ECG signals. Reliability is expressed quantitatively using a quality index (QI) for each reference heart rate. The physiological waveforms were evaluated using an SVM classifier, and the heart rate was computed using an adaptive peak identification technique that cleaned any noise produced by motion.

As part of a pioneering study investigating whether PPG devices were feasible for HRV monitoring and whether movement affected PPG reading quality and accuracy, Gil et al. concluded that PPG devices did not differ statistically significantly from the ECG reference device, and there was a strong positive correlation between them [91]. In [94], Renevey et al. propose a comfortable and easy-to-use wrist-worn device scheme that relies on the use of PPG sensors for the estimation of R-R intervals and HRV analysis during sleep. Their study showed that R-R interval measurements from wrist devices were in agreement with the ECG measured by the polysomnograph.

PPG and ECG recordings were taken before and after exercise on eight healthy subjects by Lin et al. [95]. In their study, the PPG-derived HRV closely matched the HRV derived from ECG signals. Moreover, the authors showed that in order to analyze frequency-domain features of the HRV, at least three minutes of cardiac biosignal recordings were necessary, and the HRV power spectrum distribution for three-minute data was similar to that for five-minute data. According to their report, the correlations between ECG and PPG-based HRV were found to be acceptable for individuals at rest and decreased after they performed physical exercises.

Binsch et al. studied heart rate and step count measurements of three different PPG wrist-worn wearables and compared their measurements with the ground truth [96]. According to their findings, wearable PPG wristbands could provide reliable heart rate measurements while the subjects were in idle and resting states. In contrast, the sensor readings become less accurate when the body moves during more active tasks. Ge et al. examined the accuracy of heart rate readings of two commercially available wristbands, i.e. an Apple watch and a Polar chest strap equipped with PPG and ECG sensors, respectively. These devices were used to measure heart rate data from 50 healthy participants. Their experiment was conducted in three stages, sitting for two minutes as a state of resting, walking with a speed of two km/h for two minutes as a mild physical exercise, and finally, jogging for two minutes with a speed of four km/h for the simulation of a situation with a more severe cardio activity [97].

They found that in normal conditions, when subjects were at rest without any physical activity, the results were almost identical, with a maximum error of around 2%. Nevertheless, during higher intensity physical exercises like walking, there was a difference of around 10% between the two devices. Overall, chest bands that use ECG tend to be more accurate than Apple Watch's PPG for heart rate monitoring during physical activity.

A comparison study was conducted by Ollander et al. on both time and frequency domain HRV features computed from the Empatica E4 and a reference ECG device during a Trier Social Stress Test (TSST) stress induction session. The IBI obtained from the Empatica E4 wristband displayed a substantial degree of degradation, especially when the individual was instructed to carry out a task. Despite this, time-domain features of HRV, which are widely utilized in stress detection experiments, are accurate enough to be used [98].

A recent study carried out by Mejía-Mejía et al. compared the accuracy and quality of HRV measurements using ECG and PPG devices. They administered a whole-body cold exposure while collecting PPG measurements from various body locations. They applied Bland-Altman, and analysis of variance to demonstrate that PPG not only responds differently to cold exposure in comparison to ECG but also responds differently across different body parts [93].

3.3.2. Qualitative Comparison Studies

Qualitative analysis techniques have been commonly used alongside quantitative methods of analysis in existing research [99]. Qualitative analysis involves transcribing study participants' views, opinions, or experiences. It is possible to conduct structured or semi-structured interviews. The interviewer follows a set of predetermined questions in structured interviews, while in semi-structured interviews, the interviewer is free to probe beyond those questions. Qualitative data are typically evaluated using thematic analysis [100], involving two approaches. A typical analysis approach

involves coding the qualitative data and generating themes. During coding, texts from the document are highlighted, labeled to describe their content, and then merged to generate different themes. Thematic analysis can be performed inductively or deductively. An inductive approach focuses on data to generate the themes, whereby a deductive method refers to identifying preconceived themes based on existing theories and knowledge [101]. Various software tools are available for qualitative data analysis, which support coding and theme development [102, 103]. Key examples of research on affective biofeedback interfaces using qualitative analysis include Affective Diary for identifying bodily experiences [104], Affective Health for managing stress [105, 106], and Affective Chronometry for reflecting on and regulating affective experiences [77].

4. METHODOLOGY

The purpose of this chapter is to provide an overview of the research methodologies used in this thesis. This includes the steps taken to design and implement efficient mechanisms for detecting stress and emotions and the exploratory study of designing effective intervention mechanisms to help users regulate their emotions and reduce their psychological stress.

Since the ultimate goal of stress detection and coping mechanisms is their deployment for the general public, in the studies carried out to complete this thesis, we tried to place the system's end users in the center of the design as much as possible. Accordingly, in many stages of the design methodology, it was the needs and expectations of end users and their final satisfaction that were given special priority. These steps include choosing the proper wearable devices to record the physiological signals, making the artificial intelligence in the stress detection mechanism interpretable, and finally, choosing and adjusting haptic feedback. This must be noted that we will comprehensively explain all the details mentioned above in the following sections.

4.1. Mixed-Methods Research

Existing research suggests that by combining qualitative and quantitative methods, more profound and broader information can be obtained, which cannot be obtained when using a single approach alone [107, 108]. The combination of qualitative and quantitative data and methods in a research study is known as mixed-methods research. The integration is often accomplished with the help of a team composed of quantitative and qualitative researchers.

4.1.1. Subjective and Objective Data

The term "subjective" refers to someone's personal opinion or feelings about a subject. Subjective views or opinions are not based on truth or fact. In other words, they are one individual's personal interpretation of an idea, affected by their feelings, thoughts, and background. The level of pain and discomfort experienced by a patient and their description of their symptoms are examples of subjective data. On the other hand, information that is "objective" is based on factual and data-driven information. Even though personal opinions and sentiments are not objective, objective data like facts or statistical information can form the basis for subjective feelings and ideas. Empirical and indisputable data and shreds of evidence are used to formulate an objective assessment of a subject.

4.1.2. Qualitative and Quantitative Data

Qualitative data and the process of analyzing these data, which is referred to as qualitative research, relies on data collected by the researchers coming from the original sources or personal experiences, learned or gained directly [109]. Qualitative data are usually non-numerical and can be of multiple forms, such as questionnaires (with descriptive responses), participants' subjective observations, and interviews. Research using qualitative data enables researchers to explore in-depth questions about areas of interest that are difficult to quantify. It also provides researchers with insights into how users interact with a particular technology and what their practices and experiences are [109, 110]. Contrary to quantitative data, research outcomes from qualitative data cannot be easily generalized to a broader population, as they are limited to the subjective interpretations of the participants. This issue especially becomes critical when the number of subjects is small. Quantitative data includes statistical data, percentages, and other numerical data. They are generally produced by mathematical models and experiments conducted in the laboratory or real-life settings. The purpose of quantitative research is usually to uncover patterns and relationships between data points and to test hypotheses about the interrelationships between them [111]. For this, statistics

are used to analyze the data and to produce an unbiased result that can be generalized to a larger group of people. More specifically, descriptive and inferential statistics are utilized to make sense of data features in a coherent manner and form predictions based on the data on hand, respectively [112].

4.2. Surveys and Questionnaires for Subjective Measurement of Stress

Measures of stress can be obtained from a variety of clinical subjective tests. The purpose of these tests is to collect subjective data from subjects by utilizing questionnaires. These questionnaires generally revolve around perceptions of stress and its frequency, taking into account a variety of contexts and scenarios. Almost all such surveys use Likert scales to collect users' subjective responses. Below are the questionnaires that we have used in our studies. It is worth noting that these are among the most popular and widely used questionnaires used in the literature.

4.2.1. NASA Task Load Index (Nasa-TLX)

NASA-TLX is originally categorized into two sections: In the first part, users are asked to rate six subjective sub-scales within a 0-100 range. These sub-scales are Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. The second part of the NASA-TLX is designed to establish individual weightings of the sub-scales mentioned above. In this part, users are asked to choose the most relevant measurement by comparing the sub-scales from the last part in a pairwise manner. For this, the user must choose which sub-scales are most relevant for that particular workload. Descriptions for every sub-scale are provided before the questions to help participants understand the motive and respond accurately (see Figure 4.1). Since some of the questions in the second part are not relevant to psychological stress, for instance, the problem of physical workload becomes insignificant in the identification of mental stress, in studies in which there is only psychological stress, we use a modified version of Nasa-TLX which does not include non-necessary questions.

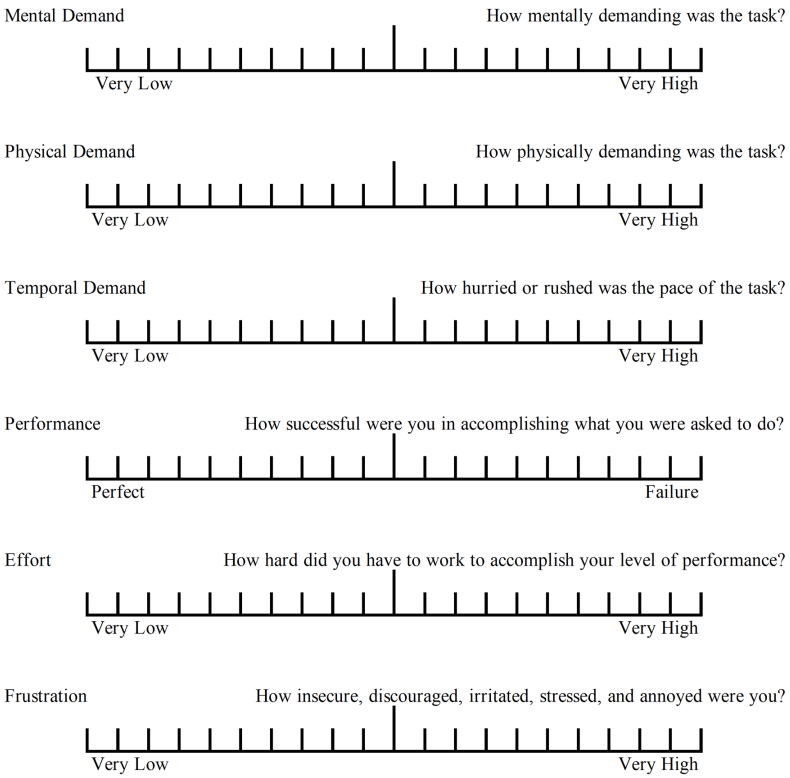
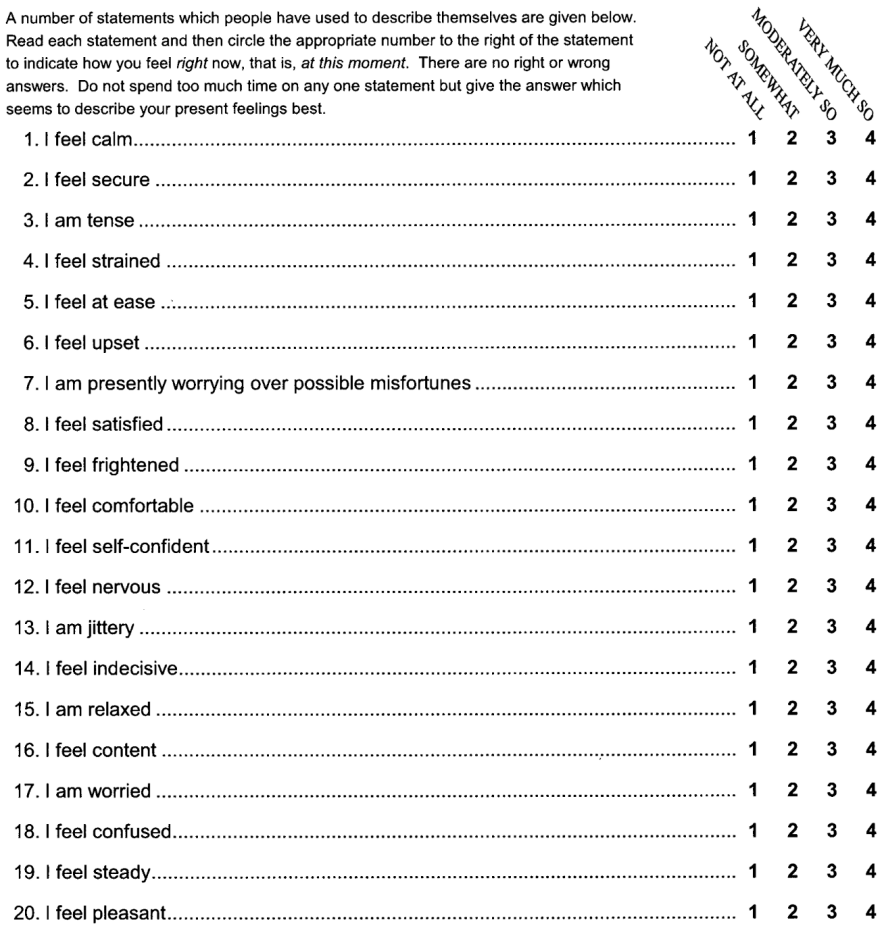


Figure 4.1. Nasa task load index (NASA-TLX).

4.2.2. The State-Trait Anxiety Inventory (STAI)

The State-Trait Anxiety Inventory (STAI) refers to a psychological survey composed of 40 self-report questions, which are on a 4-point Likert scale. There are two types of anxiety measured by the STAI, state anxiety and trait anxiety. The higher the scores, the greater the level of stress and anxiety. It is available in more than 40 different languages, and its latest version is called STAI-Y [113]. Anxiety can be characterized by feelings of stress, as well as worry and tension [114]. This test is usually administered to adults, and it evaluates how strong a person's subjective perceptions of stress and anxiety are. Various situations perceived as potentially dangerous can arouse the autonomic nervous system in "states of anxiety", such as fear, discomfort, and nervousness. As a result of perceived threats, this type of anxiety is more about how a person feels at the onset of the perceived threat and is generally viewed as a temporary condition [115].

Alternatively, "Trait anxiety" is usually interpreted as how individuals feel during typical situations such as discomfort, worry, and stress that every individual may encounter and experience on a daily basis [116]. Our studies have utilized the Y-1 version of the STAI questionnaire, which consists of 20 items to measure state anxiety, which is a momentary emotional response resulting from situations such as examinations, cognitive challenges, and especially stressful tasks such as TSST and SCWT (see Figure 4.2). A four-point Likert scale was used to rate each response, where high STAI-Y-1 scores indicated higher stress levels.



A number of statements which people have used to describe themselves are given below. Read each statement and then circle the appropriate number to the right of the statement to indicate how you feel *right* now, that is, *at this moment*. There are no right or wrong answers. Do not spend too much time on any one statement but give the answer which seems to describe your present feelings best.

	NOT AT ALL	SOMEWHAT	MODERATELY SO	VERY MUCH SO
1. I feel calm	1	2	3	4
2. I feel secure	1	2	3	4
3. I am tense	1	2	3	4
4. I feel strained	1	2	3	4
5. I feel at ease	1	2	3	4
6. I feel upset	1	2	3	4
7. I am presently worrying over possible misfortunes	1	2	3	4
8. I feel satisfied	1	2	3	4
9. I feel frightened	1	2	3	4
10. I feel comfortable	1	2	3	4
11. I feel self-confident	1	2	3	4
12. I feel nervous	1	2	3	4
13. I am jittery	1	2	3	4
14. I feel indecisive	1	2	3	4
15. I am relaxed	1	2	3	4
16. I feel content	1	2	3	4
17. I am worried	1	2	3	4
18. I feel confused	1	2	3	4
19. I feel steady	1	2	3	4
20. I feel pleasant	1	2	3	4

Figure 4.2. The state-trait anxiety inventory, version Y-1 (STAI-Y1).

4.2.3. Perceived Stress Scale (PSS)

Perceived Stress Scale (PSS) measures the degree to which a person perceives a situation as stressful. It is one of the most extensively employed psychological questionnaires for measuring overall perceived stress since it was first introduced in 1983 by Cohen et al. [117, 118]. The PSS reveals both objective physiological indicators of stress and a heightened risk for health problems among individuals with higher perceived stress levels. For instance, people with higher PSS scores (which suggests chronic stress) were more prone to depression [119]. Answer alternatives are straightforward, and the items are clear and concise. The questions are also of a general nature, so they do not address any particular target population. The PSS asks about individuals' feelings and thoughts during the past month. In each question item, subjects are questioned about how often they had particular feelings.

4.3. Participants

In all the studies we have carried out in this thesis, subjects have volunteered to participate, and we have used similar methods to find the participants for our studies. For instance, for the study in [19], campus flyers and a mailing list were used to promote the experiment and recruit participants. Many interested participants contacted us through e-mail, and we were able to schedule a suitable time for them to participate in the study. In order to participate in the study, participants were instructed to follow a usual sleep routine the night before [120]. They were also instructed to avoid drinking caffeinated and alcoholic drinks and also to avoid eating two hours before the study began [121, 122].

4.4. Ethics

Before starting the data collection sessions in all of the research we have conducted, the ethics review board of Boğaziçi University approved the studies, and we ensured that the procedure of the methodologies used in our studies complied with

the 1964 Declaration of Helsinki [123]. Each subject signed an informed consent form before data collection began. Following that, they received an information sheet explaining the study procedures. Additionally, each participant was informed that their participation is voluntary and that they are free to opt out of the study at any time during the data collection until seven days after its completion. They were told that by doing so, all data relating to them would be destroyed after the completion of the data collection sessions. An anonymous identification number was assigned to each participant, and all relationships between the participants' names, subject numbers, and data were removed.

4.5. Data Collection

4.5.1. Laboratory settings

During the data acquisition sessions in laboratory settings, all subjects were subjected to identical one-on-one sessions during a one-time visit to our laboratory. Any participant who reported serious mental or physical health issues, such as anxiety, depression, hypertension, or cardiovascular disorders, was excluded from the study. In all of our studies, participants were provided with biosensors at the beginning of the study, along with instructions on how to wear them. The experiments were conducted in the laboratory in a completely silent room without any visual or auditory distractions. We assisted the participants with wearing the sensors. Participants were instructed to take a seat that was positioned in front of a table facing the wall. Behind the participants, on another table, there were also two notebook computers that were being used for data collection. Participants were able to get used to the surrounding environment during this process, which took about 15-20 minutes.

During measurements of the vagal tone and psychophysiological processes, it is recommended that body movements be avoided to obtain clear and accurate readings [72]. Generally, baseline recordings are conducted while the subject is seated, and it is recommended that the body position be as close to the baseline as possible during

stressor and recovery sessions [72]. Therefore, in our studies, we followed the same protocol for data acquisition. We instructed participants to sit calmly with their hands on the table and avoid making any large movements following the sensor attachment. In data collection, we (researchers) did not take turns, and the study setup was the same for all participants.

4.5.2. Data Labeling

After the data collection stage, labeling them is one of the most important and challenging steps for training efficient machine learning models [124]. Depending on the nature of the data and the subjects under study, data labeling can be done during data collection or after its completion. The difficulty of labeling also depends on the type of data. For example, in medical studies related to medical imaging, the image data taken by an MRI machine needs to be labeled by an expert (a radiologist) [125].

The idea of having an expert from related scientific backgrounds, such as psychology in our case, label the data brings the ground truth closer to reality. It increases the value of the collected data set, the analysis performed on that data, and the final results obtained. However, we believe that the absence of an expert for labeling did not affect our data quality for the reasons explained below. First of all, in the data collected in the studies related to this thesis, which have been conducted in both laboratory and daily life settings, it is almost impossible for an expert to be constantly present in the daily life of each user and to monitor them constantly. Secondly, in the data collected in the laboratory environment, wholly standard and identical procedures have been applied to induce mental and physical stress. We believe utilizing these widely used standard procedures makes the study not need to be interpreted by experts. It must be noted that similar studies in the literature as well have not employed experts for data labeling when using the same standard methods for stress induction. In the following, the standard methods of stress induction used in this thesis are explained.

4.6. Stress induction

4.6.1. Psychological Stressors

4.6.1.1. Trier Social Stress Test (TSST). Trier Social Stress Test (TSST) is a well-established, widely used, and reliable method to induce stress in subjects. This method is designed for laboratory settings and consists of a combination of several procedures that have previously been known to induce stress. However, previous procedures did not seem to be able to accomplish this reliably. Clemens Kirschbaum and colleagues created this method in 1993 at the University of Trier [126].

The TSST has been implemented in a variety of ways (for instance, the original version was slightly longer), yet most current versions follow a similar pattern. The stress induction procedure lasts about 15 minutes and is composed of three components, each lasting five minutes. During the test, either an intravenous (IV) line for the purpose of collecting blood or a heart rate monitor for collecting cardiovascular signals is attached to the subject. In order to induce stress, the participant is taken to a room where three judges are waiting for the participant to begin the procedure. The room is usually equipped with a video camera or audio recorder [126]. In the first five minutes, the subject is asked to give a five-minute presentation. In the course of the test, the judges maintain a calm and non-judgmental expression. The judges observe the subject without commenting during the five minutes of the presentation component. In the event that the participant does not use all five minutes, they will be asked to continue, and this process continues until all five minutes have been completed. A mental arithmetic component follows the presentation, during which the subject is asked to count backward, for instance, from 707 in steps of 14. This part lasts for five minutes, followed by a recovery phase. Following the test, participants are informed that the test was designed to create stress and that the results are not indicative of their mathematics and presentation abilities.

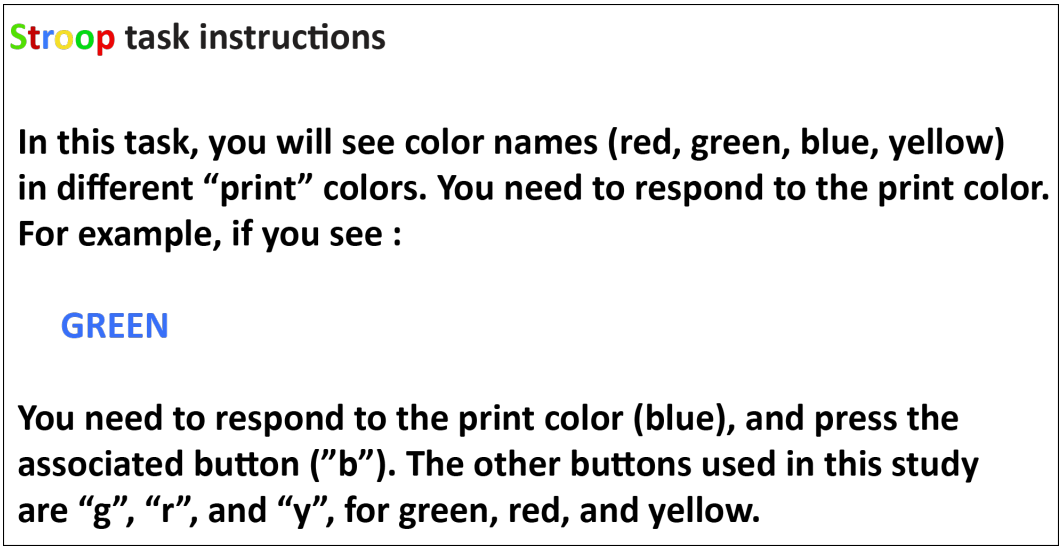


Figure 4.3. Stroop color and word test.

4.6.1.2. STROOP Color and Word Test (SCWT). The Stroop Color Test is a color and word neuropsychological test widely utilized in the literature and clinical purposes [127]. The Stroop Color and Word Test (SCWT) is based on the principle that humankind can read words more quickly than they can identify and recognize colors. During the Stroop Color Test (see Figure 4.3), the names of the colors (e.g., green, red, or blue) appear in different colors (e.g., the word "blue" in green, instead of in blue). The subject is asked to identify the color words displayed in a discrepant color (for example, the term "Green" is shown by a blue color). The processing of the distinct feature (word) impedes the simultaneous processing of the second feature (color), taking longer time and effort, making the subjects susceptible to more misinterpretation. SCWT has been widely exploited in the literature as a mental stressor [127, 128].

4.6.2. Physical Stressors

4.6.2.1. Cycling. The reason for choosing cycling as the physical activity in some of our studies is that due to limitations in laboratory environments, moderate to intense physical activity is only possible using equipment such as a stationary bike or a treadmill. HRV represents the activities of both SNS and PNS components. When an

individual engages in exercise and physical activity, it affects their HRV, as physical demands are met by both components [129]. HRV has been used in the literature for measuring physical stress [130–132] with cycling as the exercise method or even psychological stress during the cycling [133]. Generally, the Cycling activities in our studies [19], lasted for 5 minutes, where participants started with low resistance (60W) and then gradually moved to medium (90W) and ended up performing intense cycling exercise (120W). After the cycling activity, subjects underwent a five-minute recovery period, and data collection stopped after the last recovery.

4.7. Third-party Tools for Signal Analysis

4.7.1. Kubios HRV

In order to preprocess the data obtained from cardiovascular biosensors, we employed an HRV analysis tool named "Kubios" [134]. Kubios HRV is scientifically validated and is one of the most widely utilized robust HRV analysis software in the research community. Kubios can be used solely for analyzing HRV data or even for measuring stress's impact on human health using its built-in algorithms.

4.7.1.1. RR Detection. With Kubios HRV software, QRS values are accurately detected from ECG signals and pulse waves from PPG signals. Kubios detects the R peaks for any raw ECG signal using a Pan-Tompkins-based QRS detection algorithm [135]. In the studies we have conducted, only the BITalino (r)evolution board saved cardiovascular data as raw ECG, which needs to be converted to HRV using an HRV analysis software. The sampling rate for this device was set to 1000 Hz. Although sampling at this frequency is sufficient for HRV analysis, in order to provide even more accurate detection accuracies, the R peaks in the QRS detection algorithm are interpolated at a sampling rate of 2000 Hz. If ECG devices that sample at a lower rate are used, this technique enhances the temporal resolution of R peaks even more. A matched filtering technique is used by Kubios for the detection of the pulse waves from the raw PPG signals. A maximum of first derivatives is used to predict the initial pulse position.

This first derivative corresponds to the steepest part of the pulse. In the next stage, a matched filter is built using the correlation of the first pulse located in the earlier stage as the template to detect the existence and locations of a matching template (pulses) in the approaching signal parts. In most of our studies, we utilized Kubios to extract the IBI values from devices that record PPG data in its raw form, such as the Empatica E4 wristband. As for the remainder of the devices, the IBI values are automatically calculated inside the device, which can be downloaded as their output file.

4.7.1.2. HRV Artifact Removal. As we mentioned earlier in this section, data must be preprocessed before they can be analyzed. In the preprocessing stage, artifact removal is one of the most critical steps. Any signal, including biomedical or psychophysiological, is susceptible to noise and artifacts. Cardiovascular and PPG signal artifacts needed for HRV analysis usually result from poor data quality due to subjects' involuntary physical movement or environmental factors during data acquisition. They can affect various sensors differently based on the sensor kind and the location where they are attached. These factors can cause more negative impacts while the HRV is being collected in ambulatory and real-life conditions. PPG sensors, for instance, are highly susceptible to noise, and multiple factors can negatively affect their signal-to-noise ratio. It is, therefore, essential to employ reliable methods for detecting and removing artifacts from all HRV signals, especially those from wristwatch-recorded PPG signals. In order to minimize the possibility of severe deformities in HRV analysis that can result from artifacts in RR time series, the task force of the European Society of Cardiology recommends that all artifacts must be either corrected or removed [10]. In several cases of our studies, we have applied two different forms of artifact correction algorithms [19], [39]. The first one, also referred to as the "automatic correction" in the Kubios HRV software, identifies and corrects the artifacts detected on the data of a time series signal based on a procedure proposed in [136]. There are also threshold values in Kubios HRV software referred to as Very Low, Low, Medium, Strong, and Very strong for the 0.45s, 0.35s, 0.25s, 0.15s, and 0.05s threshold values, respectively. While the application of automatic correction was sufficient for the ECG devices in our studies, such as the Firstbeat Bodyguard 2, Zephyr HxM, and Polar H10, for PPG

devices, there was a need to apply threshold-based artifact corrections with medium and strong settings. This was due to the fact that the PPG devices are more sensitive to noise and artifacts and have lower data quality compared to ECG devices.

4.7.2. cvxEDA and NeuroKit2

As described in Subsection 2.1.3, skin sweat gland activity increases following the occurrence of high arousal in the sympathetic branch of the autonomic nervous system (ANS). This increased sweat gland activity increases the skin conductance. Emotional arousal and stress cause the body to sweat, which increases skin conductance [137]. Therefore, skin conductance acts as a measure of sympathetic and emotional responses. As a result, EDA is a promising candidate for detecting stress levels.

4.7.2.1. Artifact Correction and Feature Extraction. SC (Skin Conductance) signals are contaminated by intense physical activity and temperature fluctuations. Filtering out affected segments (artifacts) from the original signal is, therefore, necessary. Using an EDA toolkit, in the Skin Conductance signal (SC), we were able to detect the artifacts with 95% accuracy [138]. The artifacts were manually labeled by technicians during the development of this tool. By using the labels, the machine learning model was trained. Additionally, skin temperature and 3D acceleration signals were employed for artifact detection. From our signals, we discarded the pieces that Kubios recognized as artifacts. We further enhanced this tool by adding batch processing and segmentation. In addition, we have also utilized NeuroKit2, a Python toolbox designed for neurophysiological biosignal processing [139] in some of our studies.

A feature extraction phase followed the phase of removing artifacts from EDA signals. The signal consists of two components, phasic and tonic; features were extracted from both components (see Table 4.1). The signal was decomposed into these components using the cvxEDA tool [140]. Based on Bayesian statistics, this tool uses convex optimization to estimate the activity of the Autonomic Nervous System (ANS).

Table 4.1. EDA features and their definitions.

Feature	Description
Quartdev Tonic	Quartile deviation (75 percentile–25 percentile) of the phasic component
Strong Peaks Phasic	The number of strong peak per 100s
Peaks Phasic	The number of peaks per 100s
Perc20	20th percentile of the phasic component
Perc80	80th percentile of the phasic component
Mean Tonic	Mean of the phasic component
SD Tonic	Standard deviation of phasic component

4.8. Data Analysis

This section provides descriptions of the quantitative and qualitative data and the methods and tools utilized to analyze these data. These analyses consist of data preprocessing for HRV, and EDA signals, followed by qualitative analysis of the questionnaires and interview data related to participants' perceived data.

4.8.1. Preprocessing

4.8.1.1. Windowing. According to [10], short-term HRV recordings are recommended to last five minutes. Other researchers have studied even shorter R-R intervals, called ultrashort-term recordings (three minutes, two minutes, one minute, 30 seconds) [141]. They concluded that ultra-short-term analysis of HRV can become a new alternative to the standard five-minute analysis [141–143]. Before preprocessing the quantitative data, raw cardiovascular and EDA data must be separated into the experienced sessions, i.e., the Baseline, Event, and Recovery, based on annotations of time and context information acquired and taken note of throughout the data recording sessions.

4.8.1.2. Signal Synchronization. In case using multiple wearables for the purpose of multimodality or while conducting a comparative study, signal synchronization becomes one of the critical steps. This procedure is particularly critical when making a

comparison of signals of the identical or even similar kind captured from various devices attached to a single individual to ensure that there are no significant time lags and drifts between all of the recordings of each session. In our studies, we performed the data synchronization step automatically and manually to guarantee the most accurate data alignment. One of the widely utilized approaches in signal synchronization is cross-correlation [144–146]. We employed cross-correlation to find and rectify any potential time shift between the signals from multiple devices while taking one of the devices as the reference point. In this procedure, the cross-correlation maximum corresponds to the time-point in which the signals are most adequately synchronized. Furthermore, we tried inspecting and synchronizing the data by visually examining the HRV signals in Kubios and manually aligning the raw peaks in Kubios's signal data browser window. This approach is also utilized in the literature and discovered to be as accurate as the traditional widely accepted cross-correlation procedure [147, 148].

4.9. Wearables and Biosensors

This section briefly introduces the wearables used to collect data in our studies. It should be noted that the degrees of wearability of these commercially available devices are different, and some of them may not even be used continuously in everyday life due to their obtrusiveness. However, since their size and physics are portable and wearable to some extent (regardless of the level of acceptance by users), we will introduce them in this section. Detailed comparisons of these devices and user reviews are available in Chapters 7 and 8. Table 4.2 shows a list of wearables used in our studies and a summary of the technical specifications of each.

4.9.1. Single Sensor Wearables for HRV

- Firstbeat Bodyguard 2 serves as one of the ECG devices in several of our experiments. This lightweight wearable sensor for measuring cardiac signals and R-R intervals is validated in [177] and has been used in many studies in recent years. Once the Firstbeat Bodyguard 2 is connected to the skin, it begins recording

Table 4.2. Heart monitoring sensors used in this thesis, their placements, technical details, and a list of studies conducted using these devices.

Article	Device	Sensors	Sampling rate	Placement	Connectivity	Realtime streaming	Cloud storage	Actuator/ Display	Price
[149–152]	Empatica E4	PPG, EDA, ACC, IR Thermopile	64 Hz	Wrist	Bluetooth	✓	✓	✗	$1,690
[153–155]	Samsung Gear S2	PPG, ACC, Barometer, Gyro	100 Hz	Wrist	Bluetooth	✓	✗	Display	$149
[156–159]	Firstbeat Bodyguard 2	ECG, ACC	1000 Hz	Chest	USB	✗	✗	✗	$330
[160–163]	BITalino (r)evolution	ECG, EEG, EDA, EMG, ACC	1000 Hz	Chest	Bluetooth	✓	✗	Buzzer, Led	$190
[164–168]	Polar H10	PPG	130 Hz	Chest	Bluetooth	✓	✓	✗	$75
[169–172]	Zephyr HxM	PPG	250 Hz	Chest	Bluetooth	✓	✗	✗	$55
[173–176]	CorSense	PPG	500 Hz	Finger	Bluetooth	✓	✓	✗	$165

the data automatically. The ECG signals are processed inside the device with a sampling rate of 1000 Hz. The RR data are captured as offline data that can be accessed later via a USB connection.

- The Polar H10 chest strap is an ECG chest strap that can provide an accurate heart rate measurement at a frequency of 130 Hz. Using the Polar H10, the RR data can be recorded in real-time on a smartphone and saved in cloud storage as well.
- Zephyr HxM is another ECG device employed in our studies. It is very similar to Polar H10 in performance and almost identical in appearance and aesthetics. It can only transmit its data to the computer using a live Bluetooth connection.

4.9.2. Multisensor Wearables

- The BITalino (r)evolution board kit is a board kit produced by PLUX Wireless Biosignals. It includes multiple types of sensors and actuators and measures the ECG at a speed of 1000 Hz. Acquisition of the ECG signal in the BITalino kit is performed live via Bluetooth connection using a computer.
- With an average battery life of 32 hours with a single charge, and a charging time of fewer than 120 minutes, the Empatica E4 is capable of holding up to 60 hours of recorded data. Furthermore, it supports real-time data transfer via Bluetooth in addition to a USB connection. Considering the Empatica E4 is de-

signed exclusively for research, its continuous PPG recording capabilities are significantly improved over traditional smartwatches. Empatica E4 provides Blood Volume Pressure (BVP), Skin Temperature (ST), Electrodermal Activity (EDA), Interbeat Interval (IBI), and 3D Acceleration data through its set of integrated onboard sensors. A 64 Hz sampling rate is used to record the raw BVP signal using its PPG sensor [178].

- The Samsung Gear S2 is a commercially available and widely used smartwatch that can produce the IBI with the help of its PPG sensors. We developed an application for Samsung's Tizen OS, which allows the selection of sensors to be employed for acquiring IBI data from the Samsung Gear S2. In continuous recording mode, the Gear 2's battery can last no longer than three hours. It is also equipped with an accelerometer sensor.
- The Polar OH1 is another device used only in one of our studies. This arm-worn device monitors heart rate using a PPG sensor. In [174], its accuracy for heart rate monitoring is validated, and the results indicate reasonable agreement with the reference device. There are no options for extracting R-R interval or raw PPG data in Polar's OH1. As a result, HRV analysis is not possible with the device. Consequently, it is not subjected to quantitative analysis and is only included in qualitative analyses involving usability and acceptance by users explained in Chapter 7.
- CorSense is designed to detect heart rate signals from the fingertip of the user and provides live biofeedback for training. Using a PPG sensor at 500 Hz, it measures heart rate variability.

The details mentioned above were only a part of the on-paper specifications of these devices advertised by the manufacturers. As we proceed through the chapters, we will share the hands-on experience gained with all devices, with complete comparisons of most of the devices mentioned above in Chapters 7, and 8.

5. HOW TO RELAX IN STRESSFUL SITUATIONS: A SMART STRESS REDUCTION SYSTEM

We implemented a scheme for stress detection using physiological sensor data, which was integrated with a physical activity sensor to detect the context information. Using this mechanism, either a traditional or application-based stress management method is suggested to the user in response to the detection of high stress levels. Additionally, we compare the physiological effects of both methods on 15 international early-stage researchers (ESRs) from the AffecTech project during their eight days of training at Bogazici University. Established by the European Commission, AffecTech was a program funded by Horizon 2020 (H2020). During this training event, 15 Empatica E4 smartbands were employed to record 1440 hours (equal to 60 days) of physiological data. In order to alleviate the participants' stress levels, emotion regulation approaches based on James Gross' model [47] were put into practice (see Figure 2.2). We believe that this is the first study to suggest stress reduction approaches with regard to context information. A system like this can be utilized in real-time biofeedback applications to help detect stress levels offline. Individuals could benefit from feedback pertaining to high stress levels and instructions for relaxation strategies by employing our stress level detection algorithm in a real-world setting. An individual may also benefit from additional continuous monitoring in order to better understand the efficiency of their stress reduction practice. However, our stress detection mechanism relies on smartbands in order to be functional in everyday life.

As further explained in Section 2.4.3.1, ideally, it is the individual's context that determines the most appropriate emotion regulation solution. It would be beneficial for society if there were a system that monitored stress levels, analyzed the context of individuals, and offered appropriate relaxation methods during times of high stress. Mobile as well as traditional emotion regulation techniques, should be applied in stressful real-life situations, and their effectiveness should be compared by analyzing physiological signals.

Table 5.1. Comparison of our work with the literature studies utilizing different forms of meditation methods for stress regulation.

Article	Yoga	Mindfulness	Mobile Relaxation	Device	Signal	Suitable for Daily-life
Ahani et al. [86]	✗	✓	✗	Biosemi	EEG and Respiration	✗
Mason et al. [88]	✓	✗	✗	Digital Plethysmograph (PortaPres)	Virtual Blood Pressure Respiration	✗
Svetlov et al. [53]	✗	✗	✓	Several	HRV, EDA, sAA and EEG	✗
Puranik et al. [54]	✓	✗	✗	MPU 6050+Piezoelectric Belt+Pulse Sensor	Heart Rate + Respiration EEG	✗
Karydis et al. [87]	✗	✓	✗	Muse Headband	EEG	✗
Cheng et al. [55]	✗	✗	✓	Emotiv wireless headset	EEG	✗
Ingle et al. [89]	✗	✗	✓	8-channel Enobio EEG + piezoelectric belt	EEG + Respiratory	✗
This work	✓	✓	✓	Empatica E4 wristband	PPG, EDA, ACC, ST	✓

Studies that compare the performance and effectiveness of these methods in real-life situations are not found in the literature (see Table 5.1). Ideally, these methods must be implemented with unobtrusive wearables so that they can be worn in daily life as a wearable biofeedback system. Many people are hesitant to use a system that contains cables, electrodes, and boards on a daily basis. With such systems, comparing different states would not be possible in daily life. By using algorithms that can run on unobtrusive devices, traditional and mobile emotion regulation practices can be suggested and evaluated. An ideally designed solution must be able to detect stress, recommend relaxation methods, and monitor compliance using unobtrusive devices. We propose a system architecture and algorithm suitable for embedding in such daily life applications utilizing physiological signals such as skin temperature (ST), heart rate variability (HRV), Electrodermal Activity (EDA), and accelerometer (ACC). Here in this chapter, we present the findings of a pilot study in which our system was put to the test for regular daily activities, stress alleviation activities, as well as an event that would be stressful.

5.1. Unobtrusive Mechanism for Detecting Stress Using Smartbands

Through the implementation of our stress detection system explained in [18], users will be able to monitor their stress levels during their daily activities in an unobtrusive manner. Wearing a smartband is the only requirement for using this system. A total of

15 Empatica E4 wristbands were used in this section, where subjects wore each on their non-dominant hand. It must be noted that detailed descriptions of Empatica E4 can be accessed in Section 4.9. After proper detection and handling of the physiological signals artifacts (see Subsection 4.7.1.2), features were extracted from the sensory data and passed to the machine learning model for classification. The models were trained using feature vectors and class labels collected from the data. EDA preprocessing, artifact removal, and feature extraction were conducted in accordance with the details described in Section 4.7.2. Additionally, accelerometer and body temperature data have also been used in the research conducted in this chapter. Our system uses accelerometer sensor data for two different purposes. We started by identifying stress levels from features derived from the accelerometer. EDAExplorer Tool also uses this sensor to clean the EDA signal. EDAExplorer Tool uses a skin temperature signal to detect artifacts in EDA signals [138]. Once we divided our data into segments, multiple modalities were merged into a single feature vector.

5.1.1. Relaxation Method Suggestion Based on Physical Activity Context

The term context refers to a wide range of information which may include calendars, types of activities, locations, and intensities of activities. It is possible to infer contextual information from physical activity intensity. A lower physical activity intensity could be observed in environments with fewer restrictions, such as offices, classrooms, and public transportation, while a higher intensity could be observed outdoors. Thus, individuals will require different relaxation methods depending on their context. We used the EDAExplorer tool to calculate physical activity intensity. This is accomplished using the stillness metric and expressing an individual's stillness or motionlessness as a percentage. In order to count as still, the total acceleration must be less than a threshold (default being 0.1) for 95% of a minute. It is then possible to calculate the ratio of still minutes in a session [138]. According to the ratio of still minutes in a session, sessions below 20% were considered still, while sessions above 20% were considered active, with relaxation methods recommended accordingly. A diagram of the entire system is shown in Figure 5.1.

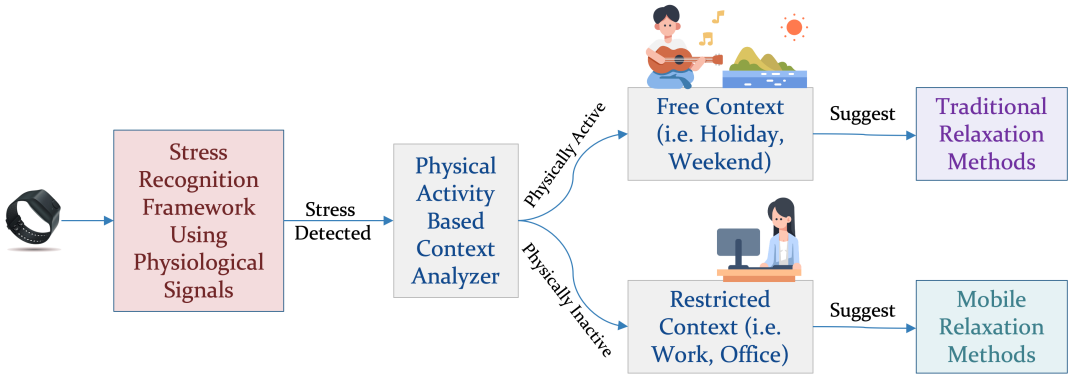


Figure 5.1. By analyzing the physical activity context, the system suggests the most appropriate method for reducing stress when a high level of stress is experienced.

5.2. An Overview of the Data Collection Procedure

The proposed mechanism for monitoring stress levels in real-world settings was fully evaluated during an eight-day training event for the AffecTech project in Istanbul, Turkey. The AffecTech project was an international collaborative research network involving 15 PhD students (Early Stage Researchers (ESR) with the aim of developing low-cost, effective wearable technologies for individuals who suffer from affective disorders. During the training event, ESRs participated in workshops, lectures, and training with clearly defined tasks and activities to ensure that they had developed the necessary skills, knowledge, and values. At the end of the eight-day training, ESRs were expected to give a presentation about their PhD research to two panel members from H2020, where they received feedback about their progress. A certified instructor conducted yoga, guided mindfulness, and mobile-based mindfulness sessions to study the effects of emotion regulation on stress. During the training, objective physiological data and subjective questionnaires (Nasa-TLX, See Section 4.2) were collected from the 16 subjects (15 ESRs and one of the AffecTech project academics). All participants gave informed consent to participate in the study as described in Section 4.4. One Empatica E4 device malfunctioned, preventing data from being included from one participant. As for the remaining 15 participants, all stages of the study were successfully completed. Figure 5.2 shows the timeline of events.

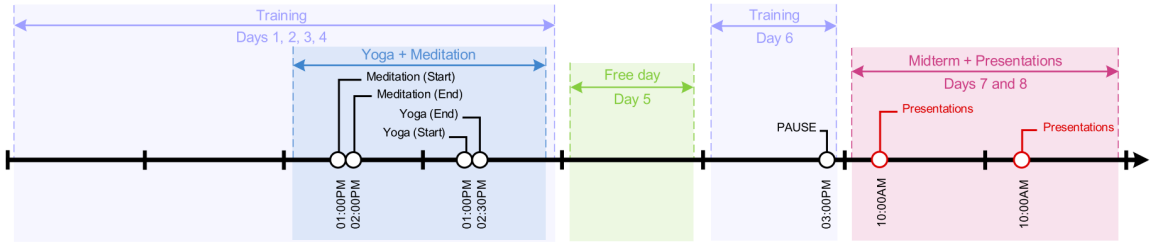


Figure 5.2. An overview of the training event over eight days. Lectures, presentations, and relaxations are highlighted.

5.2.1. Physiological Stress Data

Recorded physiological data included IBI, EDA, ACC, and ST, all saved in multiple CSV files. Additionally, 27.39% of the data comes from free time (free day and after training until subjects went to sleep at 17:00 –10:00), 43.83% from lectures in training, 11.41% from presentations, and 17.35% from relaxing sessions. In order to overcome the problem of class imbalance, we randomly undersampled the data. In order to demonstrate whether the participants' stress levels were modified prior to and after each stress reduction event (yoga and mindfulness), participants' blood pressure (BP) was also measured using a Medical-grade sphygmomanometer by a technician. Each time the participants' blood pressure was recorded, the mean of three measurements was used as the final value. Reduced blood pressure and/or pulse rate may indicate reduced stress levels, as evidenced by lower blood pressure and/or pulse rate. Since the study of the relationship between blood pressure and human psychological states is out of the scope of this thesis, we will not cover it in this chapter.

5.2.2. A Yoga and Mindfulness-based Stress Management Scheme

It was assumed that participants' stress levels increased over the eight-day training as they had to present to the H2020 project evaluators their progress during the PhD training (perceived as a stressful event). We offered yoga and mindfulness sessions on days three and four in order to help participants manage their stress levels.

Some of the questions surrounding the research conducted in this chapter were how far we can reduce users' stress by providing methods for emotion regulation and which of these methods will be more effective. To answer these questions, we presented two different approaches to stress management: a traditional method, i.e., yoga, and a modern approach using a mobile phone application, i.e., PAUSE, (see sections 2.4.3.1, and 2.4.3.2).

5.3. Validation of the Perceived Stress Levels Using Subjective Reports

In order to validate that the participants experienced varying perceived stress levels in three contexts (lecture, relaxation, presentation), we used the frustration scale in the Nasa-TLX questionnaire to assess perceived stress. Figure 5.3 shows how the answers were distributed. We aim to demonstrate that perceived stress levels (obtained from self-report answers) differ considerably between relaxation and presentation sessions (high stress). Therefore, we compared perceived stress self-report answers from yoga versus presentation, mindfulness versus presentation, and PAUSE (mobile mindfulness) versus presentation sessions using a t-test. For evaluating the separability of each session, paired t-tests are used. Each session tuple was subjected to the variance test; none of the sessions had equal variance. As a result, we selected unequal variance. Confidence intervals of 99.5% were used. For all tuples, the null hypothesis stating

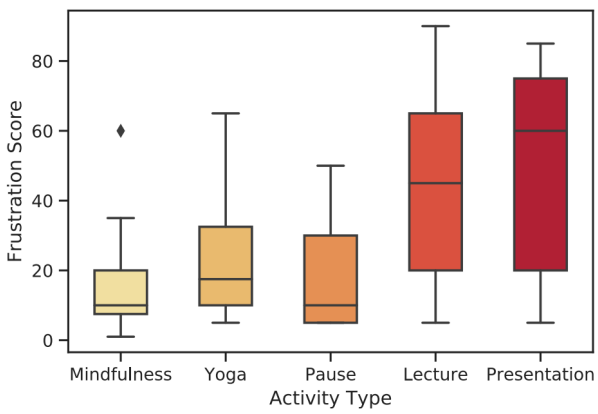


Figure 5.3. Barplots illustrating frustration scores collected in various sessions.

that the perceived stress of the relaxation method was not less than the presentation session was rejected. Participants' perceived stress levels during all meditation sessions are significantly lower than those during the presentation session (high stress).

5.4. Stress Level Detection Using Context as Class Labels

In order to test our system in this section, we used the known context of the sessions to label the classes. By examining perceived stress self-report answers, we used Lecture (mild stress), Yoga and Mindfulness (relax), and Presentation in front of judges (high stress) as class labels. It was then examined whether relaxation methods were effective, whether different modalities were beneficial, and how to find the presenter. The performance of two and three-class classification was evaluated using the interbeat-interval, skin conductance, and accelerometer signals separately and in combination. Classes include mild stress, high stress, and relaxation from mindfulness and yoga sessions. Tables 5.2, 5.3, and 5.4 illustrate the results. The most challenging part of the classification task was similar physiological responses to relaxing and mild stress scenarios. However, since our study is primarily concerned with differentiating high stress classes from other classes and offering relaxation techniques in this state, it did not have an impact on our system. Furthermore, we examined the 2-class classification performance of high-mild stress and high-relax. HRV was able to discriminate between high and mild stress with 98% accuracy when used in with Multi layer Perceptron (MLP) (see Table 5.3). In the high-relax 2-class case, only Random Forest (RF) achieved a maximum accuracy of 86% using HRV features, whereas MLP achieved a maximum of 94% accuracy with ACC features. Table 5.4 shows that the combination of all signals with RF achieved 92% accuracy, which is the best among all classifiers.

Table 5.2. System performance as a result of combining different modalities. Note that the number of classes is 3 (high stress, mild stress and relax).

Algorithm	**Accuracy (%)**			
	HRV	**EDA**	**ACC**	**Combined**
MLP	72.14	36.61	74.29	82.68
RF	67.86	36.96	86.61	85.18
kNN	65.00	29.82	70.89	78.39
LDA	69.82	31.96	73.39	85.36
SVM	47.14	30.54	58.57	46.96

Table 5.3. System performance as a result of combining different modalities. Note that the number of classes is 2 (high stress, and mild stress).

Algorithm	**Accuracy (%)**			
	HRV	**EDA**	**ACC**	**Combined**
MLP	98.00	60.00	64.00	98.00
RF	98.00	42.00	72.00	98.00
kNN	94.00	44.00	58.00	94.00
LDA	94.00	40.00	54.00	94.00
SVM	66.00	54.00	54.00	66.00

Table 5.4. System performance as a result of combining different modalities. Note that the number of classes is 2 (high stress, and relax).

Algorithm	**Accuracy (%)**			
	HRV	**EDA**	**ACC**	**Combined**
MLP	82.00	66.00	96.00	90.00
RF	86.00	60.00	94.00	92.00
kNN	82.00	66.00	88.00	90.00
LDA	78.00	64.00	92.00	88.00
SVM	78.00	62.00	52.00	74.00

As seen in Tables 5.2, and 5.3, using multi-modal data compared to data from single modalities does not always lead to better classification results. This is especially the case when one of the modalities has parts with more severe intensities than the others for some reason. For example, since there are strong differences in the amount of physical activity between different sessions in certain parts of the data, the accelerometer data leads to perfect classification results. But when we added the rest of the data that included less intense physical activity, the impact of the accelerometer data becomes lower. The fact is, models trained by partly high-intensity data, which is caused by specific environmental conditions, may not generalize well. Furthermore, regarding the results shown in Tables 5.2, and 5.3, where some models show a classification accuracy of 98%, it should be emphasized that, firstly, our data was collected from a controlled environment. Secondly, while we considered the hands-on lectures as mild stress and the final presentations as high stress, it is not difficult to distinguish between the two since the subjects had very high stress levels in the presentation session. However, their stress levels were not as high when listening to the lectures. Nevertheless, it is impossible to solve this problem so easily in real life, and it is still open to new solutions.

5.5. Effectiveness of Yoga, Mindfulness and Mobile Mindfulness

The stress levels of individuals were managed using three different relaxation methods. The effectiveness of each method was measured by how easily physiological signals in relaxation sessions could be separated from those in high stress situations. Those who perform better at classification could be inferred to be more successful in reducing stress if they can be differentiated from those with high stress levels. Table 5.5 shows that mobile mindfulness reduces stress levels less effectively than desktop mindfulness. Yoga, however, has the best classification performance and therefore is the most efficient emotion regulation method in this study.

Table 5.5. The classification accuracy of the relaxation sessions using stress management methods - (using HRV).

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

5.6. Summary and Final Thoughts

This chapter aims to detect high stress levels and suggest suitable relaxation methods (e.g., traditional or mobile) when high stress levels were experienced. We designed a stress detection framework that is unobtrusive, comfortable, and suitable for daily use. We also developed a relaxation method suggestion system that uses the physical activity context of the user to suggest relaxation methods. The majority of studies in the literature only measure individual stress levels without offering any intervention for emotion regulation. However, in this chapter, we monitored participants' stress levels and helped them manage their stress levels using yoga, mindfulness, and a mobile mindfulness application. Our results suggest that yoga and traditional mindfulness perform better than mobile application-based mindfulness.

6. PERSONALIZATION OF THERMAL AND VIBROTACTILE PATTERNS FOR EMOTION REGULATION

The purpose of this chapter is to explore the use of haptic stimuli for emotion regulation. We investigated how users engaged with the technology and explored its effectiveness for emotion regulation. In this experiment, participants created temperature and vibration-based haptic patterns. After applying a set of standard stress-inducing methods, we evaluated how these patterns affected subjects' emotion regulation.

The key objective of the study is to address the following research questions:

- How can emotion regulation be achieved using vibrotactile and thermal patterns?
- How do these haptic patterns help subjects to regulate their emotions?

The above-mentioned research questions were addressed through an exploratory research involving 23 subjects. Subjects were randomly divided into two groups. One of the groups designed either thermal or vibrotactile patterns for emotion regulation. Following that, they were given these personalized patterns via haptic actuators during a stress induction session. In the other group, the subjects only completed the stress session without experiencing any haptic feedback. We also collected self-reported subjective stress measures from all participants using STAI inventory, as well as HRV data using the Empatica E4 wearable. The State-Trait Anxiety Inventory indicated that subjects in the vibration and thermal haptic groups had significantly lower subjective levels of stress compared to participants who did not receive haptic patterns. A similar change, although not significant, was observed in subjects' HRV levels who underwent haptic patterns as well, which indicated a reduction in their stress levels, especially when exposed to vibrations with low frequency. In addition, STAI scores indicated that cold temperatures and low-frequency vibrations might have a more substantial beneficial impact on experiences of perceived stress.

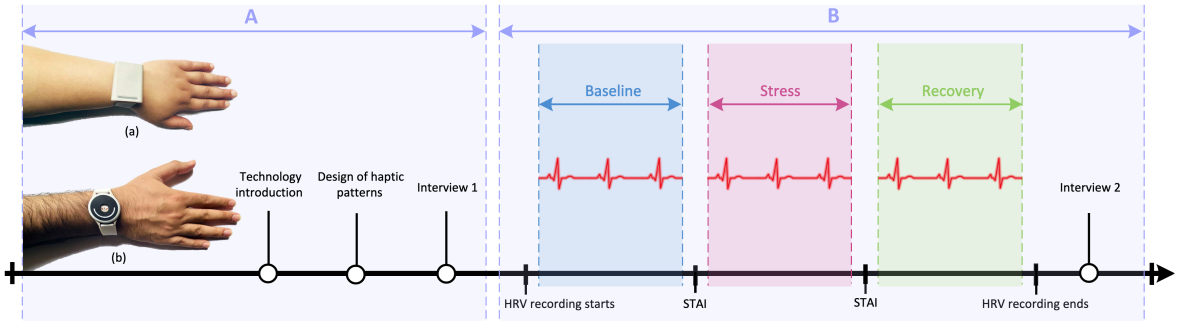


Figure 6.1. Study methodology: (a) Participants creating personalized haptic patterns (b) Stress induction procedure (both groups).

Individual sessions lasted about 60 to 70 minutes for each subject, and the method of data collection and research methodology was entirely in accordance with the method described for the studies conducted in the laboratory environment described in Subsection 4.5.1. A total of 23 volunteers were enrolled, 12 females and 11 males, with an average age of 25.4 years. Participants were randomly assigned to vibrotactile (7 subjects), thermal (8 subjects), or no haptic patterns (8 subjects). There was an equal chance that each participant would be assigned to one of the three groups. In the haptic group, participants' first step was to create their own personalized patterns to regulate their emotions. The patterns were created either through vibrotactile or thermal modalities, and the participants then utilized the corresponding actuators to experience their customized patterns during the stress induction sessions. Participants in the no-haptic group only engaged in the stress induction session in the absence of any haptic patterns and personalization. The study procedure is demonstrated in Figure 6.1.

6.1. User-Personalized Haptic Patterns for Emotion Regulation

Subjects were engaged with haptic actuators in this part to investigate and customize haptic patterns in order to regulate their emotions. As seen in Figure 6.1, in the "technology introduction" stage, the haptic interfaces were demonstrated to subjects. For this, two commercially available wrist-worn wearables depicted in Figure 6.1 were

utilized for the temperature [179], and vibration [180] haptics. A number of previous studies in the literature have used both of these actuators to produce haptic patterns based on vibration, and temperature [65], [181]. Slides were shown to the participants that described how the actuators worked and how to operate them. Only the subjects in the haptic group participated in this phase, which lasted 10 minutes. Each participant was then randomly assigned to one of the two subgroups, so they could only experiment with either vibration or temperature haptic actuators. The purpose of this decision was to limit subjects' time exposed to the stress session, prevent their physical and mental exhaustion, and keep the study duration as close as possible to 60 minutes. Following the introduction, the haptic group participated in the main exercise, which was to design haptic patterns for emotion regulation. In order to accomplish this, using the provided actuators, subjects were taught to investigate the tools and make vibrotactile or temperature-based patterns on their wrists for the purpose of calming themselves when experiencing stress. Participants explored and changed the actuators' settings via a mobile application connected to each haptic actuator via Bluetooth connection. Depending on user preferences, either type of actuator could be worn on the inside or outside of the wrist. With the vibratory actuator, subjects were able to change the frequency as well as the intensity of vibrations (from 30 bpm to 185 bpm) and (5% to 100%), respectively.

Figure 6.2 shows the settings selected by all subjects in the vibration group. Subjects were able to adjust the temperature intensity of the thermal actuators, which ranged from -11° to +16° (Celsius) from their baseline temperature, using the devices' custom-built application. By doing so, they could create thermal patterns with increases up to +16° Celsius or decreases down to -11° C. This actuator provided thermal patterns every seven seconds for the purpose of preventing heat build-up and maintaining a constant temperature, followed by another seven-second gap to avoid overheating. It should be noted that only the remaining subjects in the haptic group attended this stage, which lasted 10 minutes.

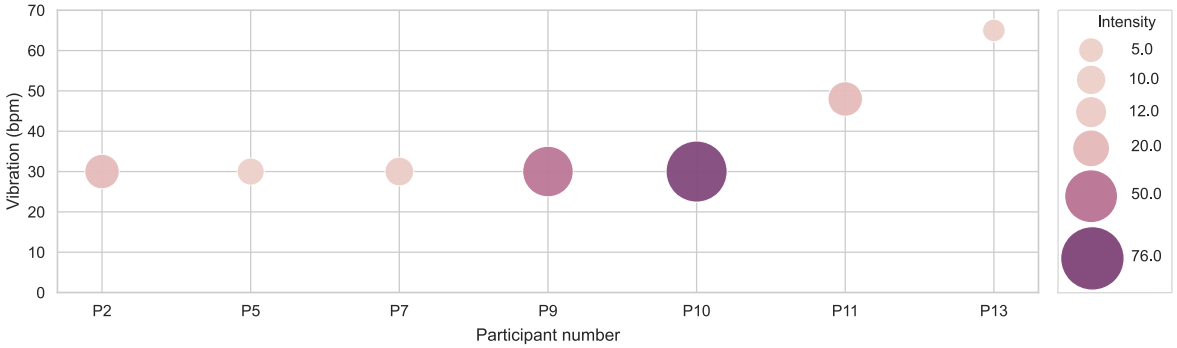


Figure 6.2. Participants create their own frequency and intensity of vibration.

6.2. Emotion Regulation Under Induced Stress: The Influence of Personalized Haptic Patterns

Following the design of haptic patterns, a stress induction session involving baselines, stress, and recovery exercises was used to evaluate their impact on emotion regulation. In the first part of the study, subjects in the haptic group were exposed to thermal or vibrotactile patterns which were designed by themselves. In contrast, no actuators or haptic patterns were provided to subjects in the no haptic group. There were three phases to this procedure. In the first phase, the stress level was measured at baseline. Neither the haptic nor no-haptic groups wore actuators at this phase. Both groups were instructed to sit comfortably for ten minutes without moving and to keep their hands on the table. After this stage was completed, the stressor task was initiated by the induction of stress using standard methods, TSST, and Stroop color test, as described in Subsection 4.6.1.2. Using two different stressors for five minutes each, stress inductions were applied to both participant groups for ten minutes. As soon as the stressor tasks were completed, haptic actuators were switched off, and a five-minute rest was given to provide the opportunity for recovery. State-Trait Anxiety Inventory (STAI) was employed to measure subjective stress. Following baseline and stressor tasks, subjects were instructed to submit their responses to the STAI questionnaires on paper. In order to estimate the objective assessment of stress, the HRV, a widely employed indicator of stress and emotion regulation, was recorded for further analysis.

HRV data has been used to extract a wide range of features using the Kubios HRV tool as described in Subsection 4.7.1. This chapter employs a widely-utilized time-domain feature of the heart rate variability, namely the RMSSD. Studies support the notion that RMSSD is closely linked with PNS activity [12], and a decrease in RMSSD indicates elevated stress, while rest and recovery are reflected in the increase in RMSSD. In order to record the HRV data, the Polar H10 sensor was used for all subjects. The Polar H10 chest strap studied in Chapter 7 records heart rate variability with a level of quality matching that of medical-grade devices. It was stated to participants that physiological signals are measured by the device, yet not specifically how and which signals. This was because there should be no bias in how haptic feedback would be perceived, such as heartbeats being associated with vibration. Findings of the study [39] presented in this section describe the effects of subjects' personalized haptic patterns on their emotion regulation by analyzing their subjective and objective stress levels. Furthermore, their perspective on using haptic technologies for emotion regulation in daily life as a result of analyzing the series of interviews conducted with the subjects was also investigated in this work [39]. Since the latter part of the study is out of the scope of this thesis, we will not cover it in detail.

6.2.1. Personalized Vibrotactile Haptic Patterns on the Wrist

6.2.1.1. Haptic Vibration Frequency. As part of their exploration and personalization, most subjects first began by changing the vibration frequency between the lowest and the highest values allowed by the device, 30 and 185 beats per minute, respectively. As depicted in Figure 6.2, most subjects preferred vibrations at a 30 bpm frequency. It is interesting to note that out of seven subjects, five of them related the vibrations to heart rate. Since an ideal heart rate would be 65 bpm when experiencing high arousal and negative emotions, these subjects created a vibration pattern with a low frequency to simulate a slow heartbeat at 30 bpm. The literature has demonstrated that vibrations with frequencies between 40 and 65 bpm can help users reduce their stress and anxiety levels [65, 66]. Subjects in our experiment chose frequencies with the same upper value and one with a lower frequency, namely 30 bpm. The lowest frequency level that our

wearable devices could support was 30 bpm, which indicates the benefit of even lower-paced vibrations for emotion regulation in stressful situations. It was reported by the subjects who selected 30 bpm that higher frequencies were more stressful, intense, and anxiety-provoking. One subject, for instance, reported feeling panicked with higher frequencies and calmed at 30 bpm. These participants were able to link the patterns of the low-frequency vibrations they experienced with their target heart rate, which is in line with the previous results in the literature [64]. Only two subjects appreciated the slow rhythms and did not relate the vibrations to heartbeats. As well as manipulating vibration intensity, subjects also modified vibrotactile sensations.

6.2.1.2. Haptic Vibration Intensity. While vibration frequency refers to the rhythm of vibration, its intensity refers to the strength of the vibration. There are two main preferences based on the findings, the majority opting for intensities below or up to 20% and the other two subjects preferring intensities above this level. Five subjects who opted for lower intensities expressed the desire to avoid intense vibrations while trying to relax and calm down. According to one of the subjects, vibrations of low intensity on their wrist were "soft". Likewise, another one stated that while a high intensity made them nervous, lower values produced a gentle sensation of touch and made them less anxious. The results suggest that it is crucial to experiment with different levels directly to discover what works for each participant. People experience physical perceptions in different ways. Research indicates that people have different perceptions of stimuli depending on their expectations, and experiences, emotions [182].

6.2.2. Personalized Thermal Patterns on the Wrist

6.2.2.1. Heat Thermal Patterns. Subjects explored different locations on and near their wrists while designing the thermal patterns. Subjects placed the actuators both on the inside and outside of their wrists. At the same time, they actuated both cool and warm with high and low temperatures were actuated. Hence, regardless of the selected same temperature being low or high, all subjects experienced the same temperature in a different way depending on the devices' placement on their wrists. Participants

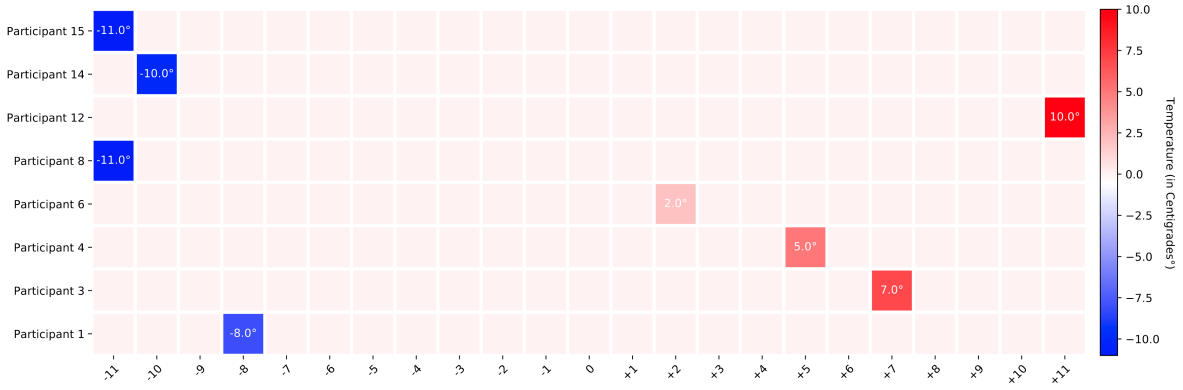


Figure 6.3. Absolute temperature values (cool/warm) in Celsius.

report preferring higher temperature intensities on the wrist's outer surface due to the inner side's greater sensitivity to temperature fluctuations. It illustrates how different parts of the wrist have varying sensitivity to temperature and the benefit of adjusting the temperature to take this into account. There is, in fact, no uniformity in thermal sensitivity across the body [183]. There are different numbers and densities of thermoreceptors across the skin, resulting in distinct nerve conduction velocities (NCV) of the cool and warm, giving rise to differences in perception [184]. According to most subjects, the inner side of the wrist that was chosen by the most number of subjects was more sensitive and easier to feel the thermal patterns than the outer side. The temperature actuator device could be operated at temperatures ranging from -11° to 16°, which subjects wore on their skin, at a standard indoor temperature (21°). According to the findings, there are two main inclinations for either increasing or decreasing the temperature. As seen in Figure 6.3, fifty percent of subjects preferred thermal patterns by raising the temperature by 7°, 5°, 2°, and 10°, whereas the remaining four preferred to bring down the temperature by -8°, -11°, -10°, and -11° (all °C). It is interesting to note that all subjects who selected higher temperatures described heat as warmth and did not exceed 10 degrees since they believed it to be quite hot. The experience of warmth was reported as comforting by all those who preferred higher degrees. The sensation of warmth was found to be relaxing and comfortable for most subjects according to a previous research on its experiential properties [70]. Warmth on the wrist was also described as comforting by the subjects in our study as well.

6.2.2.2. Cold Thermal Patterns. There were different reasons given by the other group of subjects for selecting the thermal patterns with reduced temperatures. According to them, feeling cool on a specific body part, such as on the wrist, is rather uncommon and strange since they are used to feeling heat instead of cool in daily life. Heat as a thermal modality has been explored in the literature, however, there are not sufficient studies exploring cool as a thermal modality. As reflected in their reasoning and expression, this group of subjects preferred cool temperatures over cold, comparable to the previous group that favored warmth over hot. They all found that cool felt more comfortable than warm, which is why they all selected values below -8° C. All subjects who selected lower temperature levels indicated that in addition to its pleasing nature, they prefer to experience cool since the temperature of their bodies rises when faced with a stressful situation. Therefore, the actuator's contrasting thermal patterns will assist them in noticing the situation and trying to cool down.

6.3. Impacts of Haptics for Emotion Regulation on Objective Measures of Stress

6.3.1. Between-subject Analysis

A between-subject analysis will be presented in this chapter with the aim of further investigating the effects of both types of thermal and vibrotactile haptics on the objective and subjective levels of stress. In order to limit the carryover effect and minimize subjects' exposure to the stressors, we decided to use a between-subject approach rather than a within-subject one [185]. The hypotheses were tested using the ANOVA test. ANOVA is one of the most popular statistical models used to study the differences among group means in a sample. It is a parametric approach and susceptible to errors on ordinal and not normally distributed data. We performed Shapiro-Wilk and Levene's tests on dependent variables before selecting the analysis type. While the first tests whether the sample populations are normally distributed, the latter test for assessing the equality of variances between samples. Both tests failed to prove the normality of the data and homogeneity of the variances. When these assumptions

for the one-way ANOVA are not met, ANOVA might be inappropriate and produce incorrect results [186,187]. In this case, the non-parametric methods can be used which do not rely on the population parameters, i.e., assume the necessity of the normal distribution of the data and are more appropriate for a small number of samples. We decided to conduct our statistical analysis using a non-parametric ANOVA test, i.e., Kruskal-Wallis one-way ANOVA. We used Kruskal-Wallis one-way ANOVA with intervention as an independent variable and its five levels: control, temperature and its subgroups (i.e., warm and cool), and vibration.

Kruskal-Wallis (K-W) is an omnibus test widely used in the literature to detect whether there are at least two groups among all groups that have statistically significant differences. Applying multiple pairwise K-W tests after the main effect, instead of a posthoc test, increases the amount of type-I errors, making error correction methods such as Bonferroni inevitable [188]. We executed the K-W test only once, making us in need of no Bonferroni adjustment to account for multiple uses of K-W. After finding significance in the main effect, we proceeded with posthoc tests using Dwass-Steel-Critchlow-Fligner (DSCF) [189]. It is applicable for samples of varying sizes [189] and is more suited for comparing sets with unequal variances that are also not normally distributed [190]. In order to automatically control the error rate for all comparisons, DSCF is equipped with built-in family-wise error rate protection [186], [189], [191,192]. Accordingly, DSCF does not require Bonferroni corrections. All statistical analyses in this part of the study were conducted using the Scikit Posthocs library in Python 3.9 [193], and the statistically significant conclusions were set to a significance level of 0.05.

6.3.1.1. Subjective Results. For the subjective assessment of stress and anxiety, a significant main effect of haptic patterns on participants' stress was observed via the STAI measurements ($p < 0.001$). The post-hoc analysis conducted with DSCF demonstrated that the subjects were under considerably less stress, as assessed by the STAI questionnaires' responses, with either thermal ($\mu = 35.5$, $\sigma = 12.5$, p = 0.003) or vibrotactile patterns ($\mu = 38.7$, $\sigma = 9.16$, p = 0.004), in comparison to the subjects who were given

no haptic patterns ($\mu = 56.8$, $\sigma = 3.11$). While it is important to confirm the results with future research since the number of participants was rather small, a significant finding was that the STAI scores for both haptic patterns suggest that cool temperatures ($\mu = 29.00$, $\sigma = 10.4$) and vibrations at lower frequencies (30 bpm) ($\mu = 36.4$, $\sigma = 9.04$) may potentially hold more positive impacts on the subjective perception of anxiety and stress, in comparison to warm temperatures ($\mu = 42.00$, $\sigma = 12.1$) and vibrations with higher frequencies (over 30 beats per minute) ($\mu = 44.5$, $\sigma = 9.19$).

6.3.1.2. Objective Results. For the objective assessment of stress and anxiety, the main effect of haptic patterns measured by RMSSD ($p = 0.067$) was observed to be approaching traditional significance level ($p = 0.05$): the mean values of RMSSD with the temperature haptics ($\mu = 32.4$, $\sigma = 10.7$) or with the vibration patterns ($\mu = 48.2$, $\sigma = 37.6$) were higher than the RMSSD in the absence of any haptic pattern ($\mu = 18.5$, $\sigma = 4.58$), yet again indicating lower stress level for the subjects undergoing haptic patterns. Variations in RMSSD levels under cool-warm temperature patterns and low-high frequency vibrotactile patterns were less pronounced on this objective measure of stress compared to those observed on the subjective measures. These findings demonstrate discrepancies in the levels of stress measured during the stressor and recovery tasks, respectively. Nonetheless, low-frequency vibrotactile patterns seem to be the most effective approach for emotion regulation.

6.4. Summary and Final Thoughts

This part briefly highlights the importance of users personalizing their emotion regulation patterns as a final point. Our study shows that it is indeed possible to increase the expressiveness and hedonic experience of users by entraining patterns' modality or bodily rhythm. In order to provide real-time dynamic actuation, the patterns' actuators can also be integrated with biosensors and form real-time biofeedback for emotion regulation with actuation capabilities. It must be noted that rather than being always ON and continuous, such a dynamic actuation must adjust with users' stress levels to eliminate the likelihood of overstimulation and habituation, thus

enhancing the efficiency of emotion regulation. Using thermal and vibrotactile haptic patterns (with warm/cold settings, and high/low intensity and frequency, respectively), 23 subjects explored haptic modalities for emotion regulation. These haptic patterns were evaluated for their impact on emotion regulation during stress induction tasks that were measured using self-reported stress and objective HRV features. The results showed that subjective and objective measures of stress were decreased while using haptic patterns compared to those who did not use them. It also showed that vibration actuation with lower frequency levels was the most efficient actuation type for emotion regulation. As a result of the limitations of both types of devices, which permits for changing the vibrotactile frequency/intensity, and temperature of thermal patterns solely within fixed ranges, personalization of the patterns was limited to a certain degree. However, even these narrow sets of parameters are a reasonable starting point for our exploratory research. Our findings suggest new avenues for designing emotion regulation methods for affective well-being through personalized and dynamically adjustable patterns. We believe that future work building on our research findings should look at a more comprehensive set of factors and variables. In addition, the results and findings of our study were confined to controlled laboratory environments and are thus required to be verified in real-life stressful conditions.

7. COMPARING WEARABLES FOR BIOFEEDBACK AND STRESS DETECTION USING A MIXED-METHODS STUDY

There is a growing body of work on HRV assessment using a variety of mobile, primarily unobtrusive wearable sensors. Each of these devices uses a different technology [18], and it's intended to be placed at a different location on the body [18], [93], [145]. Additionally, these devices vary in the quality of HRV data they produce and are also accepted differently by users based on wearability, aesthetics, and ease of use. When choosing a sensor for measuring HRV, the quality of the data sensing and its user acceptance are key factors to consider. A number of existing studies have compared different sensing approaches and argued for the use of either one or both [92], [194]. The majority of such studies only performed quantitative analysis to assess the correlation and agreement between methods of measurement by comparing sensing devices to a gold standard reference device [146], [149], [195]. In contrast, only a handful of studies conducted qualitative analysis for usability and acceptability [196].

As data quality and user acceptance of HRV sensing devices are equally important, this chapter extends existing studies by introducing a mixed-methods approach combining data quality and user acceptance. Combining qualitative and quantitative methods can yield greater depth and breadth of information, which is often not possible while applying a singular approach [107, 108]. We argue that combining quantitative and qualitative data would help unpack a holistic understanding of the critical factors involved in choosing a particular HRV monitoring device. This chapter explores HRV data quality concerning its features and user acceptance of different HRV wearables. In particular, we address the following research questions:

- What are the differences between the HRV measurement quality of wearable sensors and a reference device in terms of correlation and agreement levels of HRV features such as RMSSD, pNN50, Mean RR, HF, LF, and LF/HF?

- What are users' opinions of these wearables in terms of how comfortable they are to wear for daily or long-term use and how aesthetically and socially accepted they are?

Using six heart rate monitoring devices worn on key locations of the body, we combined quantitative and qualitative analyses to investigate these research questions. All of these wearables have been validated in prior research [147], [177], [197], and are widely used in the literature for HRV analysis [18], [166], [198–201]. We recruited 32 volunteers and instructed each of them to wear the six wearables at the same time. All participants were subjected to an individual data collection session, which included Baseline, Stress, and Relaxation stages, and quantitative analysis of the HRV data collected from five of these heart rate monitors were conducted using the most common and well-known agreement and correlation tests. Furthermore, in order to examine the destructive effects of artifacts on HRV data quality and how appropriate artifact removal thresholds can impact the effective recovery of noisy data, we applied three levels of artifact removal thresholds, i.e. (automatic, medium, and strong) to data collected from five HRV monitors, followed by Pearson and Spearman's correlation and Bland-Altman agreement analysis. Moreover, semi-structured interviews were conducted with participants, and thematic analysis was performed on the interview data to extract subjects' opinions and experiences regarding the sensors' wearability and comfort, long-term use, aesthetics, and social acceptance.

HRV measurement wearables are chosen based on a variety of factors. A device's efficiency and quality are influenced by several factors regardless of its application. For instance, suitability for long-term usage is always one of the sought-after characteristics. Other factors, however, are determined solely by the target application type. For most research-oriented applications, achieving results similar to the gold standard in terms of accuracy is crucial. However, such accuracy levels are not always necessary in everyday applications. Among the most desired functionalities demanded in daily life are lighter and smaller weights, ease of use along with fancy designs [196]. Simonnet and Gourvennec [202] investigated the acceptability of different heart rate

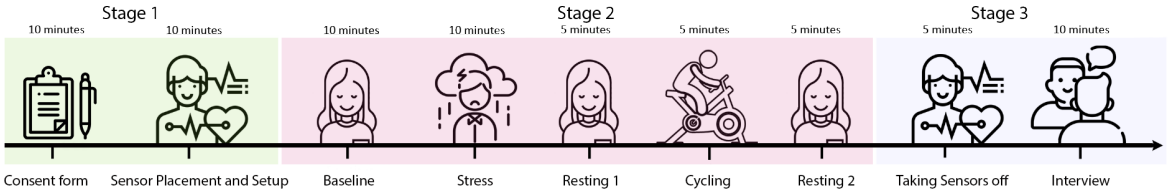


Figure 7.1. Single session individual data collection procedure for each subject.

sensors, smartwatches, chest belts, and ECG electrodes by having 11 subjects wear them for 24 hours and fill out questionnaires about the devices. These data were used to compare devices in terms of acceptability, usability, and how data reliability affects their acceptability. Accordingly, we believe that a comparison study between HRV monitoring devices should include both quantitative and qualitative factors. An extensive comparison of the six most widely used heart rate monitoring devices for multiple body locations is presented in this thesis as a contribution to existing work. Additionally, our work explores users' experiences and opinions regarding wearability and comfort, long-term usage, aesthetics, and social acceptance of HRV devices, in addition to quantitative analysis of HRV features. We conducted the study in this chapter in order to evaluate these devices so that researchers could decide on the most suitable HRV monitoring device for their research.

7.0.1. Methodology

Our study procedure comprised a 70-minute course that consisted of three main phases, as depicted in Figure 7.1. After participants consented to the study, the sensors were worn and set up in the first stage of the study. A second stage involved collecting HRV data from Baseline, Stress, Resting 1, Cycling, and Resting 2, using sensors attached to different body locations. Interviews based on subjects' observations and experiences with different wearables followed the removal of all sensors in the final stage.

7.0.1.1. Sensors. Comprehensive descriptions of the wearable biosensors utilized for capturing psychophysiological measures of stress and the software tools employed for preprocessing and analyzing these biosignals are explained in Subsections 4.9 and 4.7. For the study in this chapter, six off-the-shelf wearables were employed, including Bitalino (r)evolution kit, Firstbeat Bodyguard 2, Empatica E4, Samsung Gear S2, Polar H10, and Polar OH1. During data acquisition, we used self-adhesive silver/silver chloride (Ag/AgCl) electrodes made for medical applications to prevent artifacts mainly induced by physical movement and poor sensor-to-skin contact. Furthermore, we recommended that all participants wear the devices according to the guidelines provided to them and avoid any abrupt and unnecessary moves throughout the biosignal recording. For example, we constantly tested whether PPG devices were properly contacting participants' skin and whether they were not too loose or uncomfortably tight.

7.0.1.2. Tasks and Logged Data. In our study, we followed the same procedure described in previous studies, which involved a Baseline stage, followed by a sequence of short Stressor events followed by a Resting period [72]. In order to collect baseline data, participants sat still for ten minutes during an initial resting period. In this initial data acquisition procedure, time-synchronized heart rate data were collected for each participant. The next task involved participants completing stressor tasks for ten minutes, followed by five minutes of each of the following activities: Resting 1, Cycling, and Resting 2. There were two phases to the stressor task. In the first phase, subjects were subjected to the Stroop color and word test (SCWT) to be performed on a tablet. The SCWT was followed by the second phase of the stressor task, which consisted of arithmetic tasks, a component of the stress protocol Trier Social Stress Test (TSST) [126]. It must be noted that both types of stressors are comprehensively explained in Subsection 4.6.1. Subjects were instructed to perform backward counting for five minutes (e.g., counting backward in steps of 13 from any given 3-digit number) while one of the researchers pretended to keep track of correct and incorrect answers. Subjects were asked to engage in stationary cycling for the physical stressor part of the study, starting with low (60W resistance), medium (90W), and vigorous (120W) for five minutes. Data were continuously acquired during a five-minute resting period

following both mental and physical stressors. The collection of data was ended after the final resting stage following the cycling activity.

The last phase of the experiment consisted of a ten-minute semi-structured interview in which the subjects were asked their opinions on ease of wearing and convenience, long-term usability, visual appeal, and social acceptability of the wearable biosensors, followed by a 5-point Likert scale survey on these factors. Even though our data acquisition approach incorporates both mental, i.e., Stroop and TSST, and physical stressors, i.e., cycling, in this chapter of the thesis, we only analyze the data collected throughout the Baseline, Stress, and the first Resting phases and leave the data obtained from cycling and the second Resting sessions for the subsequent chapter. A group of 32 healthy subjects (10 Females and 22 Males, age=28.4±5.98 years, BMI=25.61±6.49) volunteered for this study.

7.1. Analyzing and interpretation of the mixed-methods data

This section presents quantitative evaluation composed of data-prepossessing, correlation, and analysis of agreement of the HRV data, as well as qualitative evalu-

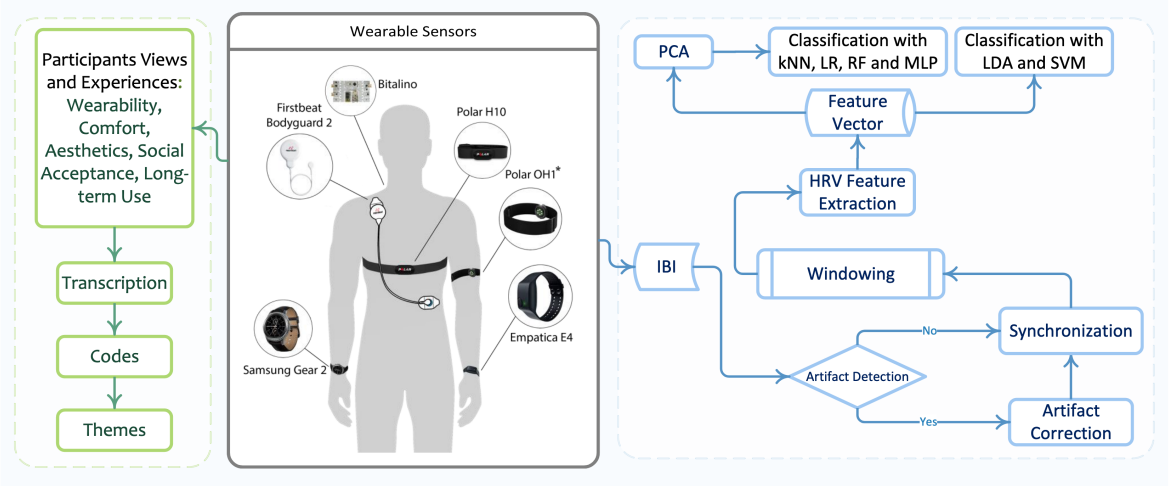


Figure 7.2. Mixed-methods approach for comparison of wearable heart rate sensors.

ation on the interview data pertaining to subjects' opinions and observations on the sensors' aesthetics, wearability, social acceptance, comfort, and long-term use as shown in Figure 7.2. Prior to conducting the final analysis of the acquired data, different preprocessing steps, including signal synchronization, RR detection, artifact removal, and feature extraction, were performed on the data. Since the majority of phases are standard procedures and thus nearly identical between several of our studies, in order to avoid repeating the explanation in each chapter, we have explained the aforementioned steps in Chapter 4. In this section, we analyzed the HRV data from short-term (5 minutes) and basic (10 minutes) recordings. Before preprocessing the data, we segmented the raw ECG and PPG data into three successive segments, i.e., Baseline, Stress, and Resting, based on timestamps recorded during the data collection, and then performed signal synchronization. Data collected from all wearables were preprocessed using the Kubios HRV analysis toolkit [134] version 3.3., described in Subsection 4.7.1.

7.1.1. Artifact Removal

For this study, we applied multiple levels of artifact removal thresholds to all data from five HRV monitoring wearables in order to investigate the destructive influences of artifacts on HRV data quality. Furthermore, we investigated how to effectively recover noisy data using reliable artifact correction mechanisms. As depicted in Figure 7.3,

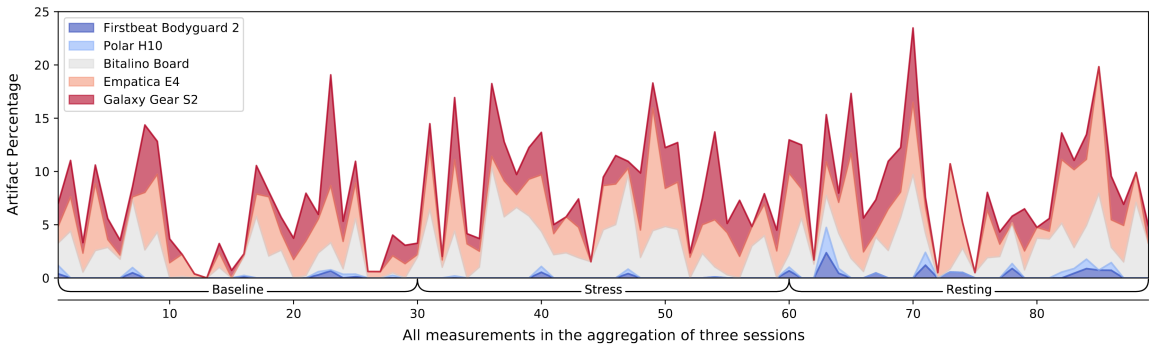


Figure 7.3. The amount of artifacts in each device during three consecutive sessions detected by the automatic correction method.

stacks of area charts are presented to visualize artifact levels of the five different devices being used concurrently by all 32 subjects throughout three successive sessions. The horizontal axis indicates the index of subjects. Due to the combination of all sessions as a whole, minimum and maximum values of x refer to the first subject in the Baseline session and the last subject in the last Relaxation session, respectively. The y-axis (in %) indicates the volume of artifacts accounted for by the automatic artifact removal procedure for each subject acquired by each wearable.

While Figure 7.3 shows the magnitude of the automatic correction approach's usage, Figure 7.4 illustrates the percentage of beats corrected by employing all artifact correction types. In order to accomplish this, we used an unsupervised clustering algorithm, K-Means, on the percentage of beats corrected in the aggregation of three sessions in each row. The number of clusters was set to equal the number of subjects (32) to visualize the aggregated sessions as summaries for all subjects. Rows were then clustered hierarchically. It is interesting to note that as a result of hierarchical clustering, the patterns of similarity between wearables and the type of applied artifact correction method are clearly evident in clusters. Only two devices belong to the same cluster: the Polar H10 and the reference device. The rest of the devices are either grouped with variants of themselves that have undergone a higher artifact correction threshold or with other equivalent families of devices.

Figure 7.4. Percentage of beats corrected for each device and hierarchical clustering for grouping the devices.

The following section compares different devices across all sessions with correlation analysis. Additionally, it presents the influence of different artifact correction thresholds on the correlation values of the results derived from different devices.

7.1.2. Correlation Analysis

Application of Correlation analysis on two sets of quantitative variables is performed to evaluate whether there is a relationship between them and what is the strength of such a relationship. While a high correlation indicates that two or more variables hold a strong relationship with each other, a weak correlation suggests that the variables are hardly related. The correlation analysis is a statistical method for analyzing the relationship between two variables and is closely linked to the linear regression analysis, which is a technique for interpreting the relationship between two quantitative variables [203]. Variables selected for correlation analysis can be two columns of any given data set. In our case, we have sets of observations recorded by multiple devices. These observations are also called samples. Each column of this sample corresponds to the values from a particular device.

There are several types of correlation coefficients, each with functionalities and usability of its own. In all correlation coefficients, the strength of the correlation is assumed to be a value ranging from -1 to 1, where -1 indicates the strongest negative correlation, 1 indicates the strongest positive correlation, and 0 indicates no correlation. The Pearson correlation coefficient is one of the most widely used techniques utilized in method comparison studies. For instance, during an investigation to validate the data quality and PPG sensor efficacy of Empatica's E4 model under various scenarios, Menghini et al. utilized Pearson product-moment correlation (r) to evaluate the strength of the linear association between the measurements from the Empatica E4 wristband, and a reference ECG device [149]. Vescio et al. used it to compare HRV measurement between the Kyto earlobe PPG sensor and an eMotion Faros ECG device [145]. In [146], Barrios et al. use Pearson's correlation as a part of their analysis to evaluate the accuracy of PPG sensors for HR and HRV measurements in the

wild. Last but not least, in [204], Lier et al. have conducted a comprehensive study on validity assessment protocols for physiological signals from wearable technology. They suggest using cross-correlation as a generalization of Pearson's correlation for the validity assessment of PPG and ECG devices at the signal level.

Despite the popularity and widespread use of Pearson's Correlation, some researchers are against employing it in method comparison studies since this method is very sensitive to non-normality in the distribution of the variables. However, we can argue that time-series data related to medical measurements such as heart rate and HRV are prone to have non-gaussian distributions. Therefore, we should use this method only to represent the strength of the linear relationship between our variables and without using the p-value for statistical significance since it assumes that data is normally distributed. So, the adverse effects of the samples that are not normally distributed would impact only the significance test rather than the correlation itself.

When the samples are not normally distributed, data has strong outliers, and the relationship between the variables is not linear, it is recommended to use the Spearman rank correlation method. The Spearman rank correlation method makes no assumptions about the distribution of the data. It evaluates monotonous relationships between two variables, whether it is linear or not, and it is equivalent to the Pearson correlation between the rank values of those two variables. In [205], Bulte et al. analyzed the association between HRV and PRV using Spearman's rank correlation coefficient for assessing the level of agreement between heart rate variability and pulse rate variability. Gilgen-Ammann et al. used Spearman to assess the correlations between the RR values from Polar H10, and an ECG holter [147] and Schrödl et al. utilized it to analyze the correlations between HRV features recorded by earlobe PPG and chest ECG [206]. There are also research cases in which both methods have been used [207].

In our study, we applied both methods to the time domain and frequency domain features of the HRV values recorded from five heart monitoring devices, namely, three ECG (Firstbeat Bodyguard 2, Polar H10, and BITalino) and two PPG sensors (Empat-

Table 7.1. Correlation values with different levels of artifact removal thresholds.

Device	Artifact Correction		Firstbeat Bodyguard 2 (Reference Device)						
	Type	Corrected	Method	RMSSD	Mean RR	pNN50	HF	LF	LF/HF
Polar H10	AUTOMATIC	0.12%	Pearson's r	1.000	1.000	1.000	0.994	0.998	0.957
			Spearman's r_s	1.000	1.000	0.998	0.999	0.999	0.998
BITalino (r)evolution		3.10%	Pearson's r	0.484	0.847	0.777	0.342	0.546	0.810
			Spearman's r_s	0.660	0.871	0.804	0.716	0.771	0.757
Empatica E4		3.21%	Pearson's r	0.521	0.993	0.936	0.227	0.321	0.672
			Spearman's r_s	0.809	0.983	0.928	0.862	0.759	0.782
Samsung Gear S2		2.58%	Pearson's r	0.412	0.993	0.759	0.362	0.540	0.589
			Spearman's r_s	0.580	0.991	0.722	0.685	0.743	0.725
BITalino (r)evolution	MEDIUM	7.58%	Pearson's r	0.667	0.832	0.850	0.524	0.865	0.616
			Spearman's r_s	0.748	0.883	0.833	0.780	0.910	0.735
Empatica E4		4.78%	Pearson's r	0.884	0.997	0.961	0.873	0.769	0.798
			Spearman's r_s	0.891	0.989	0.938	0.916	0.902	0.840
Samsung Gear S2		3.26%	Pearson's r	0.603	0.996	0.787	0.656	0.846	0.461
			Spearman's r_s	0.585	0.993	0.722	0.696	0.814	0.645
BITalino (r)evolution	STRONG	12.26%	Pearson's r	0.836	0.832	0.938	0.694	0.807	0.688
			Spearman's r_s	0.882	0.887	0.886	0.859	0.917	0.793
Empatica E4		7.77%	Pearson's r	0.904	0.997	0.971	0.860	0.821	0.830
			Spearman's r_s	0.931	0.989	0.947	0.935	0.907	0.877
Samsung Gear S2		6.94%	Pearson's r	0.720	0.995	0.831	0.749	0.792	0.526
			Spearman's r_s	0.730	0.993	0.774	0.797	0.836	0.703

ica E4 and Samsung Gear S2). Correlation analysis was performed on the aggregation of sessions to reduce the susceptibility of Pearson's correlation to the presence of strong outliers in small samples. Analysis of the correlation between the reference device and five other devices was performed with three different thresholds of the artifact correction on the devices with a higher amount of noisy data. Results of the correlation analysis for all features under study are shown in Table 7.1. The averages of corrected artifacts of all participants in the aggregation of three consecutive sessions expressed as the percentage of corrected beats are also presented in Table 7.1.

When the artifact correction was set to automatic, Polar H10 showed the highest possible correlation for all features with r values of 1 in the time domain, and r values higher than 0.95 for the frequency domain features. In the rest of the devices, although

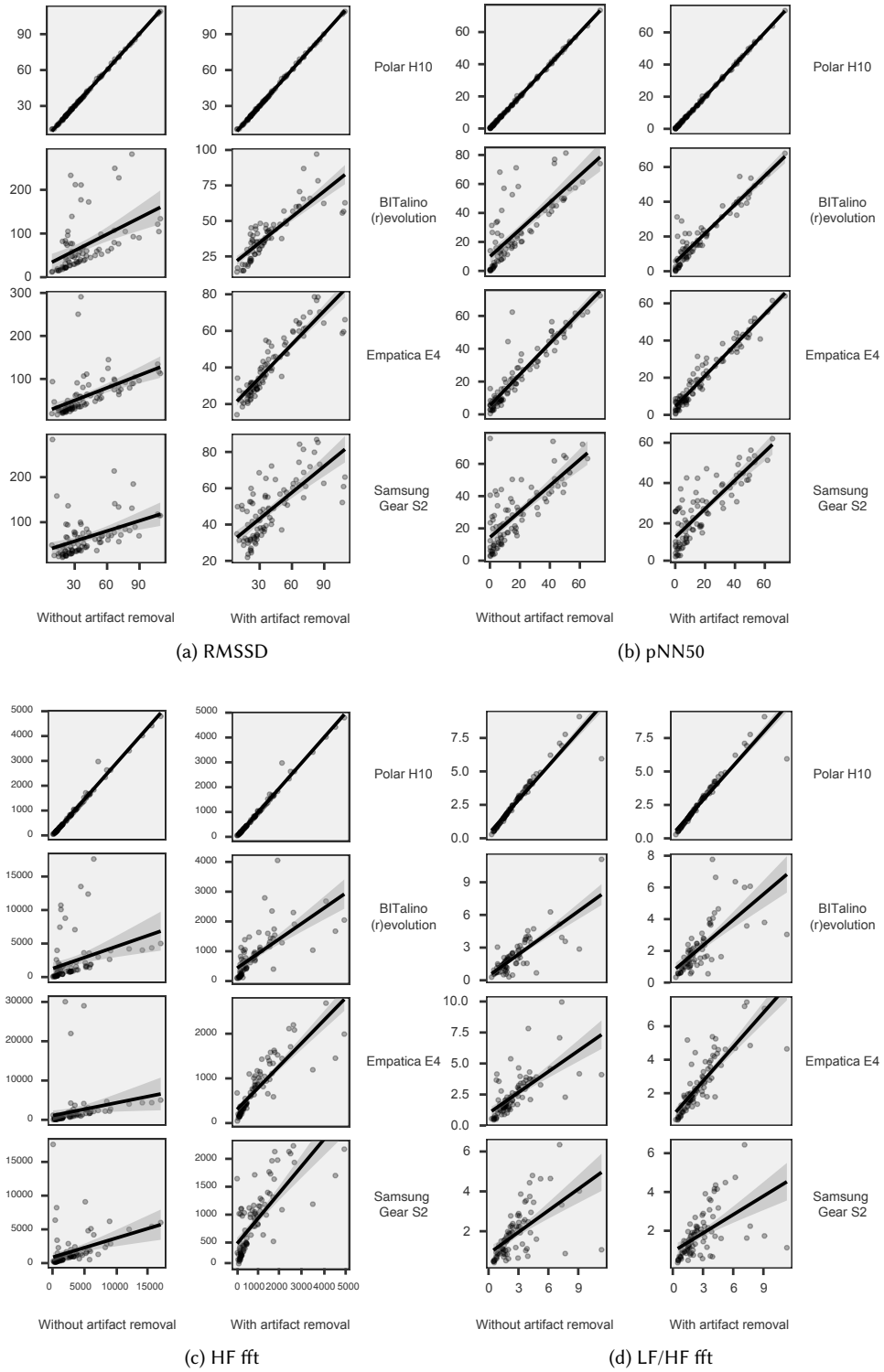


Figure 7.5. Scatter plots with linear regression line and standard error depicting the effects of artifact removal on the increase in correlation values. (a) and (b) illustrate two time-domain features. (c) and (d) represent two frequency-domain features.

there were high correlations for MeanRR and pNN50, correlation values dropped dramatically in other features, mainly in the frequency domain. It was evident from these results that further artifact corrections were necessary. In "Medium artifact Correction", we could see meaningful rises in all r values. As an instance, compared to the "Automatic" results, Pearson's r value became more than double for the Empatica E4, showing a high correlation in all features. Since the Samsung Galaxy Gear and the BITalino (r)evolution board still suffered from low correlation rates, we further intensified the artifact correction to the Strong threshold. At this level, Empatica E4 showed an even higher correlation for all features, with an average of 0.957 for the time domain and 0.837 for the frequency domain, which showed a very high correlation with the reference device. The BITalino (r)evolution board and Samsung Gear S2 also achieved higher correlations compared to previous thresholds, with averages of 0.868 and 0.848 for the time domain and 0.729 and 0.689 for the frequency domain, respectively.

Figure 7.5 shows the scatter plots with regression lines (lines of best fits) and the standard error. This line minimizes the squared difference between the line of best fit and each data point. The X-axis represents the reference device (Firstbeat Bodyguard 2), and the Y-axis represents the device being compared to the reference. These plots are similar to Pearson's correlation results, which show a positive and linear relationship between two variables. The strongest possible linear relationship occurs when the regression line's slope equals 1, and this only happens when the line of best fit lies at a 45-degree angle. By visual inspection of the plots, it is clearly visible that only Polar H10 shows a correlation almost equal to 1 in all feature types, each depicted in Figures 7.5a, b, c, and d. In other devices, the regression line moves towards 1, and error rates decrease after the application of artifact removal.

7.1.3. Bland-Altman Agreement Analysis

A clinical measurement device, in our case, heart rate monitors, must have sufficient agreements with the reference golden standard devices in the quality of measurement. Correlation and regression analysis performed in the previous section are among

the most popular and widely used methods in the literature. However, there is a debate that the correlation analysis only evaluates the relationship between one variable and another, not their differences, and it is not sufficient for thoroughly assessing the comparability between two different devices [208].

In a series of articles by J. M. Bland and D. G. Altman, an alternative analysis was proposed based on the quantification of the agreement between two quantitative measurements by studying the mean difference and constructing the limits of agreement [209–212]. The Bland-Altman plot and analysis, which is also called the limits of agreement or the mean-difference, is a comparison technique used for making a comparison between two measurements of the same variable. For instance, an obtrusive and expensive heart rate measurement system might be compared with a low-cost and unobtrusive one. The Bland-Altman (BA) method is a graphical approach in which the discrepancies between the two methods are plotted against their means [210]. Any systematic variations between the two pairwise measurements, such as fixed or proportional bias and the intensity of possible outliers, can be identified using Bland-Altman plots. While y-values depict the differences between the two measurements, the pairwise measurements' averages are assigned to the x-axis. There are also other horizontal lines parallel to the x-axis, representing the average of differences (estimated bias) and the limits of agreement. Limits of agreement (LoA) in a Bland-Altman plot are the averages of the differences ± 1.96 times its standard deviation. Computing the 95% limits of agreement (LoA) helps us understand how far apart the HRV feature measurements recorded by two different devices are more likely to be for most subjects. The Bland-Altman only depicts the ranges between the upper and lower limits of the agreement without asserting whether these limits should be considered acceptable. If a predefined a priori exists as the acceptable limits, it should be verified whether the limits of agreement on the Bland-Altman plot exceed this allowed difference or not. If the LoA are within the acceptable limits, we can state that the values inside the mean ± 1.96 standard deviation of the differences are not clinically significant, and the two measurement techniques (devices) are in agreement and can be used interchangeably. To the best of our knowledge, there is no a priori acceptable limit defined for HRV fea-

tures hitherto. Some researchers have proposed their own methods, such as accepting ±50% of the mean of the reference value as an acceptable limit or creating acceptable ratios by calculating the ratio of half the range of limits of agreement and the mean of averages. We believe that acceptable limits of accuracy and acceptability for HRV measurements must be issued from professional organizations specialized in this field. For example, the American National Standard for the advancement of medical instrumentation (ANSI/AAMI) accepts heart rate monitors as accurate if their measuring error does not exceed ±5 bpm or ±10% [213].

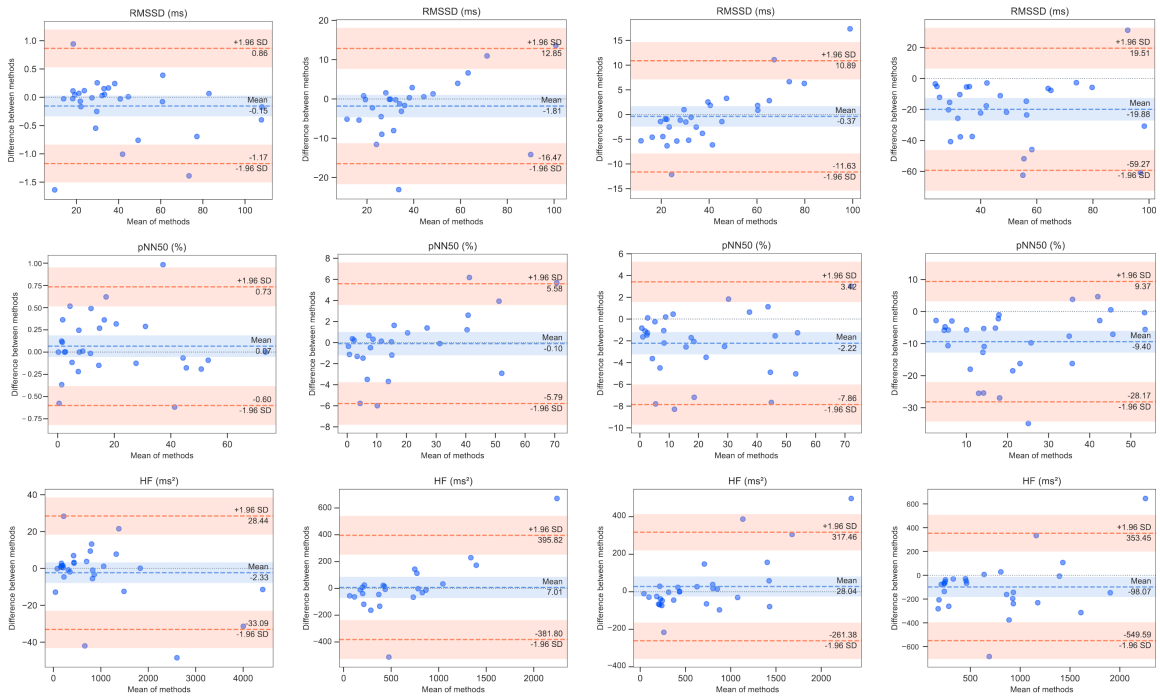
Accordingly, we interpret our Bland-Altman results based on visual inspection of the relevant plots and compare our devices against each other and to the reference device. Bland-Altman plots are represented in Figures 7.6, 7.7, and 7.8. These plots represent six HRV features in three different sessions. The artifact correction levels for the Bland-Altman analysis are automatic for the reference device and Polar H10. For the rest of the devices, this value is set to Strong. Quantitative results of the BA plots with two more features (eight in total) are depicted in Tables 7.2, 7.3, and 7.4. In most of our Bland-Altman plots, the scattered points' density is higher on the left side of the plots. Such a problem can easily be solved by performing a log transformation on the samples. However, we can reveal any possible relationship between the measurement differences of both methods and the magnitude of measurements by plotting the actual values of the HRV features. Except for the Polar H10 and the reference device, a maximum of two strong outliers were removed from the rest of the devices in order to prevent the serious negative effects of influential outliers on BA results. Since the number of devices in this study is four (five, including the Firstbeat Bodyguard 2 as the reference device), instead of showing all 96 BA plots, we decided to bring only the most important ones. However, detailed statistics of all 96 cases are presented in Tables 7.2, 7.3, and 7.4.

In the subsequent paragraphs, we will analyze the results of the Bland-Altman plots based on visual observation and comparison between the multiple wearables under each individual scenario. The Polar H10 chest band consistently shows the greatest lev-

els of agreement with the reference device across all features, and its mean bias is close to zero across all features. We anticipated obtaining similar high-quality data levels based on our experience through previous data collections using BITalino (r)evolution kit. Despite high expectations, however, it did not live up to expectations. Due to the fact that it is a board kit, it is highly susceptible to getting contaminated with artifacts when the subject moves. Since the device is not wearable, poor skin contact and attenuated signal reading can also result from incorrect sensor placement by subjects. We experienced the same issue in our study as well. Nevertheless, when we examine the correlation values and Bland-Altman levels of agreement results for the BITalino (r)evolution kit, we clearly observe the positive effects of appropriate artifact removal applied to its noisy data. Except for the Stress session in which BITalino (r)evolution displays indications of systematic errors by producing mean shifts that are significantly lower or higher than zero, it performs much better than the Samsung Gear S2 PPG wristband in time-domain features of the Baseline and Recovery sessions, especially in the normalized LF and HF features. BITalino (r)evolution's poor results in the Stress session are consistent with its greater artifact values in that session, as depicted in Figure 7.3. It is worth mentioning that in the next chapter, we will demonstrate the remarkable effects of proper noise reduction and data normalization on the stress measurement results using the data obtained from the BITalino (r)evolution kit. While the Empatica E4 exhibits proportional error in RMSSD only during the Baseline session, there is no indication of this trend in the two subsequent sessions. The Empatica E4 displays good agreement levels in the Baseline session, even with both PPG wearables exhibiting errors in the pNN50 time-domain feature. Polar H10 is the only device that does not exhibit systematic errors in the time-domain features of the Stress session. The Empatica E4 shows generally good performance for the rest of the stress and resting session features with a modest proportional error in time-domain features. Among all the devices studied, the Samsung Gear S2 offers the poorest performance. In almost all time-domain features, it displays systematic errors. As all errors in time-domain features have similar patterns, they can be adjusted. However, there are also numerous errors in the frequency domain, which makes this device the least accurate one, with the lowest level of agreement with the reference device.

Table 7.2. Bland-Altman results for the "Baseline" session.

BASELINE		Polar H10			BITalino (r)evolution			Empatica E4			Samsung Gear 2		
		95% Confidence Interval			95% Confidence Interval			95% Confidence Interval			95% Confidence Interval		
		Estimate	Lower	Upper	Estimate	Lower	Upper	Estimate	Lower	Upper	Estimate	Lower	Upper
RMSSD	Bias	-0.154	-0.348	0.0395	-1.81	-4.83	1.21	-5.76	-7.79	-3.73	-19.9	-27.53	-12.2
	Lower LoA	-1.172	-1.507	-0.8370	-16.47	-21.71	-11.24	-16.22	-19.73	-12.71	-59.3	-72.49	-46.1
	Upper LoA	0.863	0.528	1.1984	12.85	7.62	18.09	4.70	1.19	8.21	19.5	6.29	32.7
Mean RR	Bias	0.0330	-0.0920	0.158	4.67	1.58	7.77	-1.30	-4.63	2.02	0.491	-1.05	2.04
	Lower LoA	-0.6232	-0.8392	-0.407	-10.66	-16.01	-5.30	-19.09	-24.84	-13.34	-7.467	-10.14	-4.80
	Upper LoA	0.6891	0.4730	0.905	20.00	14.65	25.36	16.48	10.73	22.23	8.450	5.78	11.12
pNN50	Bias	0.0664	-0.0607	0.193	-0.103	-1.27	1.07	-2.22	-3.30	-1.15	-9.40	-12.98	-5.82
	Lower LoA	-0.6007	-0.8203	-0.381	-5.788	-7.82	-3.76	-7.86	-9.72	-6.00	-28.17	-34.35	-21.99
	Upper LoA	0.7335	0.5138	0.953	5.582	3.55	7.61	3.42	1.56	5.27	9.37	3.19	15.55
HF (ms^2)	Bias	0.389	-3.86	4.64	53.3	-203	310	12.5	-49.1	74.0	-147	-244	-51.1
	Lower LoA	-21.500	-28.85	-14.16	-1217.8	-1662	-774	-304.6	-411.0	-198.2	-644	-811	-477.6
	Upper LoA	22.278	14.93	29.62	1324.4	880	1768	329.6	223.2	436.0	349	183	516.1
LF (ms^2)	Bias	3.46	-4.33	11.2	58.7	-45.1	163	67.5	-23.7	159	62.3	-43.6	168
	Lower LoA	-36.65	-50.10	-23.2	-434.2	-614.0	-254	-393.6	-551.4	-236	-473.3	-656.5	-290
	Upper LoA	43.56	30.10	57.0	551.6	371.8	731	528.7	370.9	686	598.0	414.7	781
LF/HF	Bias	0.00960	-0.00863	0.0278	0.222	-0.0462	0.49	0.229	-0.182	0.640	0.275	-0.435	0.984
	Lower LoA	-0.08432	-0.11584	-0.0528	-1.053	-1.5182	-0.588	-1.968	-2.679	-1.258	-3.450	-4.676	-2.223
	Upper LoA	0.10353	0.07201	0.1350	1.498	1.0325	1.963	2.427	1.716	3.137	3.999	2.773	5.226

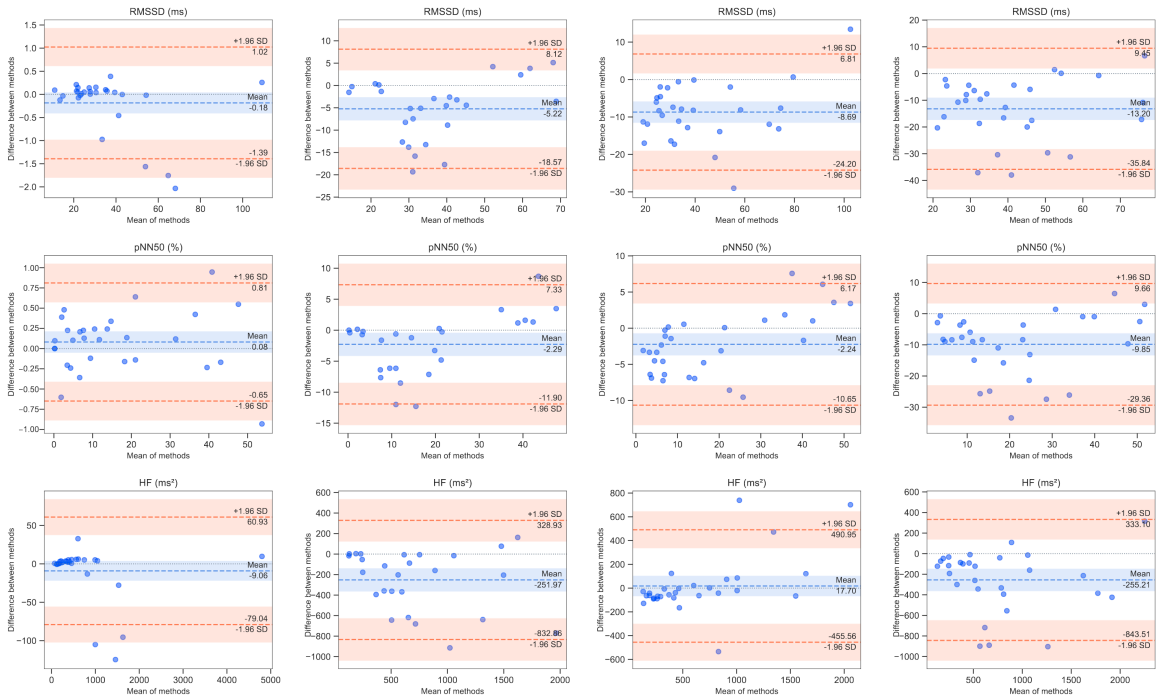


(a) Polar H10 (b) BITalino (r)evolution (c) Empatica E4 (d) Samsung Gear S2

Figure 7.6. Bland-Altman plots for the "Baseline" session.

Table 7.3. Bland-Altman results for the "Stress" session.

STRESS		Polar H10			BITalino (r)evolution			Empatica E4			Samsung Gear 2		
		95% Confidence Interval			95% Confidence Interval			95% Confidence Interval			95% Confidence Interval		
		Estimate	Lower	Upper	Estimate	Lower	Upper	Estimate	Lower	Upper	Estimate	Lower	Upper
RMSSD	Bias	-0.185	-0.424	0.0542	-5.22	-7.97	-2.47	-8.69	-11.70	-5.69	-13.20	-17.59	-8.80
	Lower LoA	-1.392	-1.805	-0.9789	-18.57	-23.33	-13.81	-24.20	-29.40	-19.00	-35.84	-43.44	-28.25
	Upper LoA	1.023	0.610	1.4359	8.12	3.36	12.88	6.81	1.61	12.01	9.45	1.85	17.05
Mean RR	Bias	0.0877	-0.0769	0.252	35.8	8.57	63.0	-2.61	-4.30	-0.920	-1.98	-4.70	0.735
	Lower LoA	-0.7761	-1.0605	-0.492	-99.1	-146.22	-52.0	-11.49	-14.42	-8.569	-15.99	-20.69	-11.289
	Upper LoA	0.9515	0.6671	1.236	170.7	123.57	217.8	6.27	3.35	9.195	12.02	7.32	16.723
pNN50	Bias	0.0811	-0.0580	0.220	-2.29	-4.27	-0.306	-2.24	-3.84	-0.633	-9.85	-13.56	-6.13
	Lower LoA	-0.6492	-0.8897	-0.409	-11.90	-15.33	-8.470	-10.65	-13.41	-7.876	-29.36	-35.78	-22.93
	Upper LoA	0.8114	0.5710	1.052	7.33	3.90	10.756	6.17	3.41	8.943	9.66	3.24	16.09
HF (ms^2)	Bias	-2.23	-10.1	5.60	-259	-379	-139	-130	-245	-14.1	-518	-747	-289
	Lower LoA	-41.80	-55.3	-28.26	-841	-1048	-633	-725	-925	-525.6	-1674	-2070	-1278
	Upper LoA	37.34	23.8	50.88	322	115	530	466	266	665.9	638	243	1034
LF (ms^2)	Bias	-0.0773	-6.73	6.57	-131	-296	34.4	42.2	-59.0	143	24.7	-109	159
	Lower LoA	-33.6827	-45.18	-22.18	-916	-1203	-629.7	-469.2	-644.2	-294	-665.8	-897	-434
	Upper LoA	33.5281	22.03	45.03	654	368	940.8	553.5	378.6	729	715.3	484	947
LF/HF	Bias	-0.00339	-0.0253	0.0185	0.557	0.222	0.893	-0.262	-0.688	0.164	0.881	0.549	1.213
	Lower LoA	-0.11845	-0.1563	-0.0806	-1.106	-1.686	-0.525	-2.539	-3.275	-1.803	-0.860	-1.433	-0.287
	Upper LoA	0.11167	0.0738	0.1496	2.220	1.639	2.801	2.014	1.278	2.750	2.622	2.049	3.195

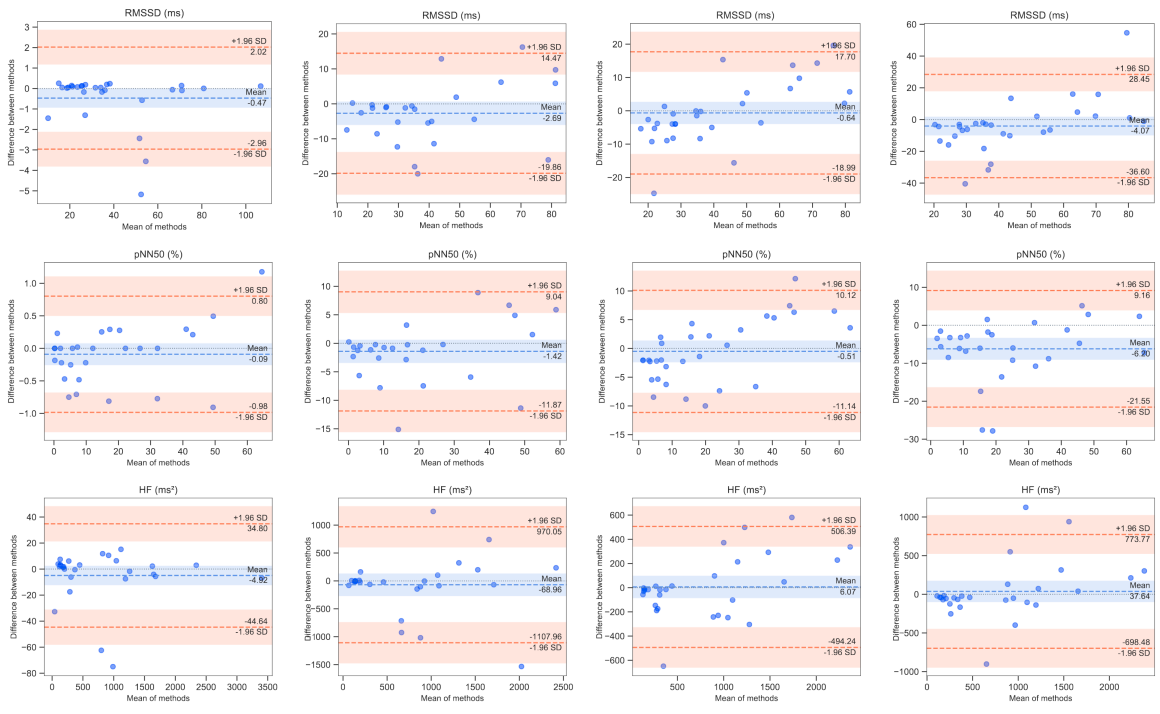


(a) Polar H10 (b) BITalino (r)evolution (c) Empatica E4 (d) Samsung Gear S2

Figure 7.7. Bland-Altman plots for the "Stress" session.

Table 7.4. Bland-Altman results for the "Resting" session.

RESTING		Polar H10			BITalino (r)evolution			Empatica E4			Samsung Gear 2		
		95% Confidence Interval			95% Confidence Interval			95% Confidence Interval			95% Confidence Interval		
		Estimate	Lower	Upper	Estimate	Lower	Upper	Estimate	Lower	Upper	Estimate	Lower	Upper
RMSSD	Bias	-0.471	-0.964	0.0226	-2.69	-6.23	0.845	-0.644	-4.14	2.85	-4.07	-10.3	2.12
	Lower LoA	-2.963	-3.816	-2.1103	-19.86	-25.98	-13.734	-18.985	-25.02	-12.95	-36.60	-47.3	-25.89
	Upper LoA	2.022	1.169	2.8751	14.47	8.35	20.599	17.698	11.66	23.74	28.45	17.7	39.16
Mean RR	Bias	-0.0618	-0.507	0.384	8.12	-1.64	17.9	-4.67	-7.56	-1.77	1.01	-3.55	5.57
	Lower LoA	-2.3571	-3.127	-1.587	-38.22	-55.13	-21.3	-19.57	-24.57	-14.57	-22.04	-29.92	-14.15
	Upper LoA	2.2336	1.463	3.004	54.46	37.55	71.4	10.24	5.23	15.24	24.06	16.18	31.95
pNN50	Bias	-0.0906	-0.267	0.0860	-1.42	-3.57	0.737	-0.508	-2.50	1.48	-6.20	-9.24	-3.16
	Lower LoA	-0.9833	-1.289	-0.6778	-11.87	-15.60	-8.141	-11.135	-14.57	-7.70	-21.55	-26.81	-16.30
	Upper LoA	0.8021	0.497	1.1076	9.04	5.31	12.766	10.120	6.68	13.56	9.16	3.90	14.41
HF	Bias	-11.4	-27.4	4.68	-96.8	-257	63.0	52.0	-57.7	162	-45.4	-191	99.9
	Lower LoA	-92.5	-120.2	-64.72	-872.7	-1149	-595.9	-513.2	-702.8	-324	-794.1	-1045	-542.9
	Upper LoA	69.7	42.0	97.49	679.0	402	955.8	617.2	427.5	807	703.3	452	954.5
LF	Bias	-2.45	-8.64	3.75	-57.2	-215	101	118	-109	345	189	-20.0	397
	Lower LoA	-33.16	-43.88	-22.43	-808.9	-1083	-535	-1072	-1463	-680	-845	-1205.4	-484
	Upper LoA	28.27	17.54	38.99	694.5	420	969	1308	916	1699	1222	860.9	1583
LF/HF	Bias	0.0472	-0.00795	0.102	0.284	-0.0510	0.619	-0.00197	-0.210	0.206	0.669	0.265	1.073
	Lower LoA	-0.2372	-0.33260	-0.142	-1.342	-1.9227	-0.762	-1.07493	-1.435	-0.715	-1.373	-2.072	-0.675
	Upper LoA	0.3317	0.23623	0.427	1.911	1.3304	2.491	1.07099	0.711	1.431	2.711	2.012	3.410



(a) Polar H10 (b) BITalino (r)evolution (c) Empatica E4 (d) Samsung Gear S2

Figure 7.8. Bland-Altman plots for the "Resting" session.

Some of the plots depict samples tending toward limits of agreement, mainly on the right side of the plot. Accordingly, we can conclude that the variation of measurement on that particular feature strongly relies on the magnitude of measurements. In addition, the concentration of samples on the left side of the plot also contributes to some of the proportional errors. It is possible to eliminate these problems in two ways: by log transformation or by shortening the window size, resulting in an increase in sample size. Furthermore, it should be noted that a consistent measurement bias can be corrected by subtracting the mean difference from the second method.

7.1.4. Users' Views and Experiences: Wearability, Comfort, Aesthetics, Social Acceptance, and Long-term Use

Descriptions of interviews conducted in the last phase of the study are presented in this section. Specifically, we describe participants' experiences regarding wearability, comfort, aesthetics, long-term use, and social acceptance of the wearable sensors. The Atlas.ti, which is a software tool for qualitative analysis of textual data, was employed for transcription of the interview data for thematic analysis [100], [102]. Atlas.ti, facilitates importing documents, marking quotes from the text, and transforming them into codes that are subsequently used in theme development. A total of over 100 codes involving participants' interpretations and experiences were generated, which were later put to use to develop themes.

Subjects referred to wrist-worn wearables such as Galaxy Gear 2, Empatica E4, and also the Polar OH1 armband as more comfortable to wear compared to other devices: *"I find wristbands easier to wear on a daily basis."* - S14 (Subject 14). Devices worn on wrists were described as comfortable and lightweight: *"These wearables were incredibly comfortable to wear."* - S3. Likewise, the Polar OH1 armband was also reported to be very comfortable, and almost all subjects stated that they did not realize they were wearing one until when it was time to remove it: *"I did not even notice this armband, that is most likely the comfiest, and easiest to put on and take off."*. However, since the electrodes of the Empatica E4 constantly press on the skin, some

participants found it quite sore and heavy after a while: *"I felt the Empatica E4 quite heavy after some time."* - S16. Unlike wrist and arm-worn devices, participants found Polar H10 chest band ECG wearable more challenging to wear since the bands must be tightly wrapped around the chest. Furthermore, subjects reported that the chest band caused discomforting pressure on the abdomen, making them feel uncomfortable when seated: *"While the chest straps are not very tight, they can make breathing difficult in a seated position."* - S5. Bodyguard 2 and BITalino (r)evolution need three self-adhesive Ag/AgCl electrodes which attach to the skin at specific locations for accurate signal acquisition. During the removal of these electrodes, all participants said they felt uncomfortable: *"It hurts when you take off the electrodes."* - S15. Furthermore, subjects reported that these devices are not convenient for daily use since they need multiple wires: *"I felt something hanging around my body when I used devices with electrodes."* - S21.

Subjects preferred the Samsung Gear S2 smartwatch over Empatica E4 smartwatches primarily for its design and aesthetics. They pointed out that in addition to providing features to measure heart rate data, smartwatches also come with various functionalities for daily use. They referred to it as *"aesthetically pleasing"*, and *"stylish"* - S7, S18, S31. In addition, subjects also stated that smartwatches are perceived as more acceptable than other wearables and will be more likely to be worn in the long term. However, chest straps were hidden underneath clothes, so they could not be regarded as socially unacceptable by other people. Subjects characterized them as being mostly sporty in nature: *"Wearing these may be limited to specific things, such as training or gauging fitness levels because they are not as comfortable."* - S10. Moreover, subjects viewed First Beat Bodyguard 2 and Bitalino (r)evolution as medical equipment. According to them, it is obtrusive, bulky, and would attract people's attention and create the impression that the wearer has a medical condition if worn in public places and over long periods of time: *"I imagine they look like the sensors patients wear in the hospitals and wearing it would give people the impression that you are wearing it for medical purposes rather than for recreational purposes."* - S23.

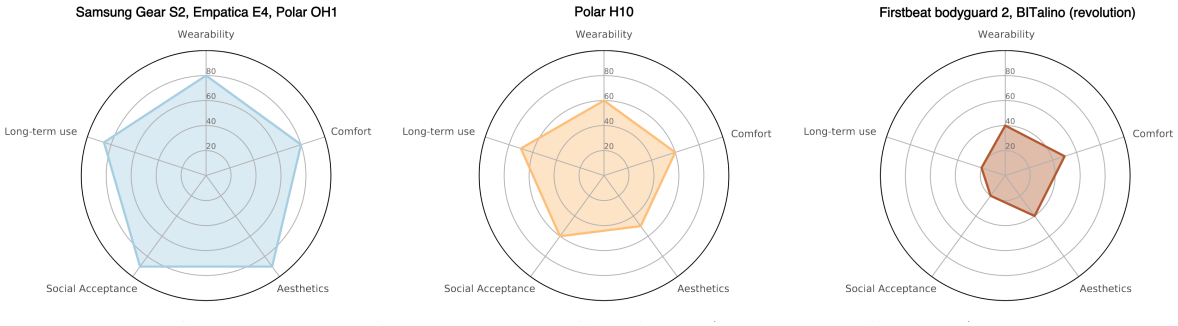


Figure 7.9. Wearability factors based on subjects' opinions and feedbacks.

A 5-point Likert scale questionnaire was administered to the subjects at the end of the interview. The results of the questionnaire are shown in Figure 7.9. Results corroborate the qualitative analysis emphasizing that subjects' choice of wrist and arm-worn devices in terms of wearability, comfort, aesthetics, long-term use, and social acceptance.

7.2. Summary and Final Thoughts

As part of our controlled laboratory study, we utilized six widely used wearable biosensors to record heart rate variability data during baseline, stress, and relaxation sessions. Five of these devices were evaluated for their data validity and quality using quantitative analysis of HRV data. The results indicate that in all sessions, Polar H10 demonstrated the most positive correlation and agreement levels with the reference device (First beat bodyguard 2), as well as having the lowest artifact levels. It was followed by the Empatica E4 wristband, BITalino (r)evolution kit, and Samsung Gear S2 smartwatch. In the agreement analysis, it was found that the wrist-worn wearables exhibited symptoms of systematic and moderate proportional errors along with much lower correlation levels when compared to the reference device and the Polar H10, in particular when it came to frequency domain features. Considering that the data acquisition was conducted in a controlled lab environment, we did not expect to see excessive amounts of artifacts in any device. In addition, all devices were worn simultaneously, ensuring that all of them were almost equally susceptible to noise con-

tamination caused by movement or environmental factors. Despite this, many artifacts were seen with BITalino (r)evolution kit, Samsung Gear S2, and the Empatica E4. How these artifacts affect the correlation and agreement analysis was investigated, and it was concluded that taking proper measures to remove artifacts, such as proper selection of artifact removal thresholds, could diminish their adverse effects to a minimum. Thematic analysis was also conducted to analyze the views and experiences of participants regarding user acceptance of wearable biosensors. Study results indicate that when it comes to aesthetics, wearability, and comfort, subjects opt for Samsung Gear S2, Empatica E4, and Polar OH1, followed by Polar H10, First beat bodyguard 2, and BITalino (r)evolution kit. In addition, subjects reported that First beat bodyguard 2, BITalino (r)evolution followed by Polar H10 are more likely to provoke negative comments from other people, which would discourage them from wearing it in public. Polar H10 was preferred for short-term use, followed by Samsung Gear S2, Empatica E4, and Polar OH1 for long-term use.

Over the past few years, there has been an increase in the utilization of mobile and unobtrusive wearable sensors for the measurement and analysis of heart rate variability. These wearables utilize ECG and PPG biosensors to collect the HRV data and incorporate lightweight and compact components that can be worn effortlessly and unobtrusively. The study in this Chapter provides additional contributions to the body of knowledge by presenting a mixed-methods design incorporating both quantitative and qualitative data analysis techniques using six HRV monitoring devices at various body locations.

In order to choose a suitable device, it is crucial to determine what type of activity you will be performing and what your predetermined goals are. Researchers working with affective biofeedback technologies for stress measurement could use the quantitative data analysis presented in our study to help them choose a sensing device that is optimal in terms of both usability and user acceptance. For use cases with low physical movements and activity, where there is no need for extreme accuracy, wrist-worn wearables i.e. Empatica E4 and Samsung Gear S2, provide a good balance

between accuracy, wearability, and comfort. For a level of accuracy that is close to medical grade, we vouch for Polar H10 and Firstbeat bodyguard 2. However, they cannot be worn for extended periods of time due to limited comfort and wearability. Multi-modal biosignal analysis can be explored with BITalino because it offers highly customizable settings and usage.

Certain limitations exist in the wearables that can lead to shortcomings in situations requiring longer HRV recordings. For example, the relatively short battery lifespan of the Samsung Gear S2 makes it impractical to record continuously for more than three hours. One additional example of a drawback is the necessity for continuous Bluetooth connectivity to a third-party mobile or computer application for data recording in devices such as BITalino (r)evolution kit, which makes such wearables less suited for recording outside the lab settings.

We obtained our study results in the laboratory settings under identical 70-minute sessions comprised of baseline, stress, and relaxation sessions, all administered while the subjects were sitting. We argue that researchers should be aware of the strengths and limitations of HRV measurement wearables prior to conducting studies. As users choose devices to monitor their heart rates and HRV, we hope the research in this Chapter will provide guidance to help them make an informed decision about the trade-off between data accuracy and usability.

8. APPLICATION LEVEL PERFORMANCE EVALUATION OF WEARABLE DEVICES FOR STRESS CLASSIFICATION WITH EXPLAINABLE AI

We evaluated the performance of different wearables for identifying mental stress and physical activity. We also employed a state-of-the-art explainable AI (XAI) method to investigate and demonstrate the importance of features and data preprocessing techniques and their impact on the output of various classification algorithms. Comparisons of different wearables were performed regarding accuracy and sensitivity of diagnosis. In addition, several factors, such as differences in the impact of features, were examined and compared, which will be explained in detail in the following sections.

Data collection and research methodology of this chapter is completely in accordance with the method described for the studies conducted in the laboratory environment described in Section 4.5.1. Seven different off-the-shelf heart rate monitoring wearable devices were utilized in this study (see Figure 8.1): BITalino (r)evolution board, Firstbeat Bodyguard 2, Polar H10, Zephyr HxM, Empatica E4, Samsung Gear S2, and CorSense. The technical details of the devices used are listed in Table 4.2.

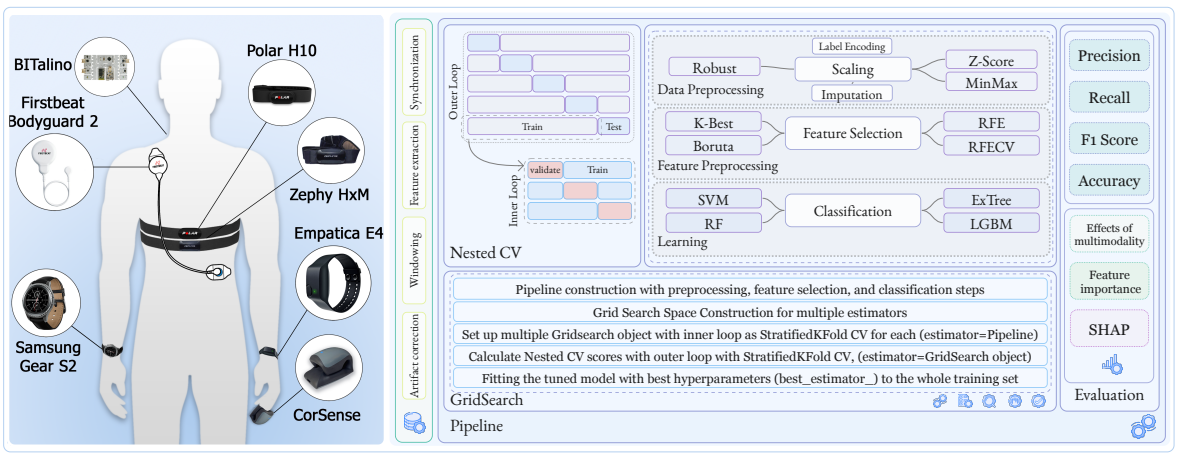


Figure 8.1. A brief outline of our system architecture demonstrating several steps.

In addition, the first column of this table represents a number of studies in the literature that have used these devices in the research related to stress detection and measurement. Moreover, since the price can be one of the important criteria for choosing a wearable device, the price of each wearable device at the time of writing this thesis is also mentioned in the last column of this table. Further information on the sensors and data acquisition are detailed in Chapter 4.

8.1. Data Preprocessing for Classification

The importance of data preparation always precedes their analysis, and any form of neglect at this point can negatively affect the study results. Since real-world data is not always perfect, it may contain strong outliers and missing values. In the data preprocessing stage, we first need to replace the missing data caused by some technical and software problems using imputation. The total amount of incomplete values in our data was negligible, and the number of missing values due to technical problems was also insignificant. We performed the imputation only in sporadic cases, where a small part of a particular session was lost for a specific device. It is a fact that the normal ranges of physiological signals of different people are not in the same exact range [214]. Since the intensities and strengths of signals differ and the accuracy with which different devices record the signals, this means that the features are not in the same ranges and do not have the same weight. Using such data in machine learning can lead to misleading results. Scaling is necessary for several ML algorithms to bring all features to the same level and ensure that a single number with its large magnitude does not negatively impact the model. Therefore, as part of data preprocessing, we need to scale the data, meaning that the data need to be transformed to fit within a specific scale. In many ML classifiers, if features were not in an approximately standard normally distributed fashion, ML algorithms would not behave well. One of the common requirements to resolve this issue is the data standardization process.

Except for classifiers based on decision trees, many classification algorithms are developed in line with the hypothesis that features must obtain values near zero and

that all feature values fluctuate on equivalent scales. If features are not presented to these algorithms as a standard normally distributed set, their prediction performance can be impaired. For example, a Support Vector Machine with an RBF kernel is not able to properly learn from feature values that exhibit much smaller variances than others, and it would be dominated only by the features with very large variances. One of the common requirements to resolve this issue is the process of standardizing the data. Using robust scalers is ideal if the data contains outliers. There are several types of scaling algorithms, each employing its own method for estimating the parameters for shifting and scaling the data. In this study, we tested several different algorithms to achieve the best results.

The StandardScaler (Z-score Standardization), MinMaxScaler (min-max normalization), and RobustScaler in scikit-learn were used in the preprocessing step. When the data is distributed in a Gaussian manner, StandardScaler may be more appropriate. It performs the standardization by subtracting the mean and then scaling to Unit Variance or dividing all the values by the standard deviation. MinMaxScaler subtracts the minimum value of the feature from each individual value and then divides the result by the range, which is the difference between the maximum and minimum values of the feature. Despite preserving the shape of the original distribution, MinMaxScaler does not alter the information embedded in the features meaningfully. However, it is vulnerable to the effects of outliers and cannot mitigate their influence. The MinMaxScaler returns values between zero and one as its default range. RobustScaler applies the same scaling principle as MinMaxScaler. However, instead of using the minimums and maximum of a feature, it uses the interquartile range, making it more robust against outliers.

8.2. Preprocessing Pipeline

The machine learning pipeline is one of the best solutions to make ML models optimized, scalable, and automated. It is a process that enables ML workflows to be automated by transforming and correlating data in a model that can later be evaluated

for results. ML pipeline is an iterative process involving several steps to train a model, in which each step is repeated in an attempt to improve the results continuously. Multiple estimators can be chained together using a pipeline. This method is helpful because there is generally a fixed sequence of operations for data preprocessing, for instance, standardization, feature selection, and classification.

Furthermore, pipelines ensure that the same samples are used to train transformers and predictors in cross-validation, thus preventing statistics from the test data from leaking into the trained model. Information leakage from test data to a trained model may lead to unrealistic and overly optimistic results far from the truth.

8.2.1. Feature Selection

Feature selection is a crucial concept that can significantly impact a model's performance by removing irrelevant features. In order to reduce the complexity and the time required for the execution of computations, which has been greatly increased due to the utilization of nested cross-validation, feature selection becomes one of the essential steps in constructing our stress detection model. Following the selection of a set of popular feature selection algorithms, in order to achieve the best results, a preliminary comparison was made between the impact of using each of them on model accuracy (See Figure 8.2).

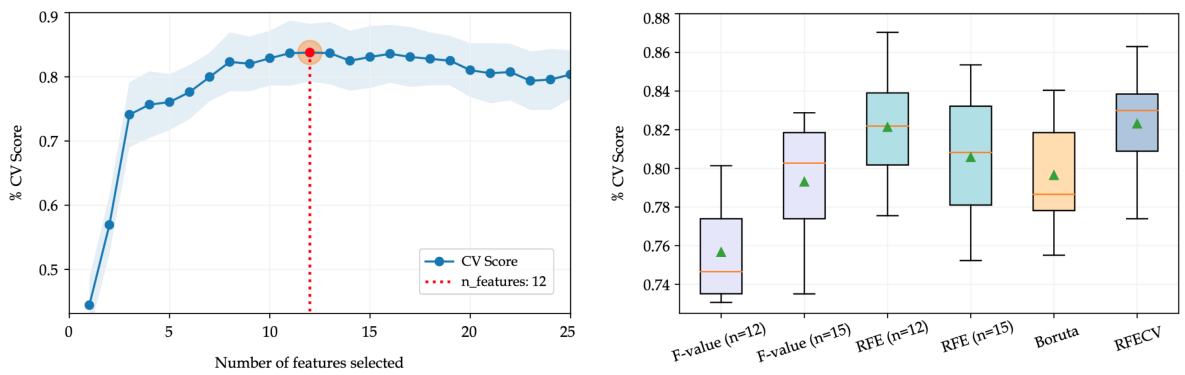
In order to make fair comparisons, it is essential that all conditions be identical, especially the quantity and the nature of the data being compared. Considering that the primary objective of this study is to compare the accuracy of stress detection in different devices using the same models, all comparison conditions should be the same for all devices. The final comparison results will not be fair if the number and type of the features selected vary between each device. Following the initial implementations of different feature selection algorithms on all training data sets, we found out that the set of features selected by the Recursive Feature Elimination with Cross-Validation (RFECV) led to the best classification results. In addition, according to RFECV, the

Table 8.1. List of the features selected by RFECV*.

Feature type	Feature name	Description
Time-domain	Mean RR *(ms)*	The Mean of RR intervals
	STD RR *(ms)*	Standard Deviation of RR intervals
	TINN *(ms)*	Baseline width of the RR interval histogram
	HR Max - HR Min	Difference of the Minimum and Maximum HR
	HRVti	The integral of the RRI Histogram divided by the Height of the Histogram
	RMSSD *(ms)*	Square root of the mean squared differences between successive RR intervals
Frequency-domain	VLF power *(log)*	Absolute powers of Very Low-Frequency Power of HRV
	HF power *(log)*	Absolute powers of High-Frequency Power of HRV
	LF/HF ratio	Ratio between LF and HF band powers
Nonlinear	SD2/SD1	Ratio between SD2 and SD1
	ApEn	Approximate Entropy
	SampEn	Sample Entropy

**Features in this table are not ordered by importance.*

number of features selected to achieve the best classification results was between 12 and 15 out of a total of 25 features for different devices, as seen in Figure 8.2a, and Table 8.1. Additionally, the list of the top 12 features selected by four different feature selection algorithms is presented in Table 8.2.



(a) (b)

Figure 8.2. (a) Number of features selected by RFECV, (b) Effects of different methods and the number of optimal features selected by each algorithm.

Table 8.2. List of the features selected by four different algorithms.

	Feature Selection Algorithms			
Order	**RFECV**	**RFE**	**Boruta [215]**	**K-Best**
1	Mean RR *(ms)*	Mean RR *(ms)*	Mean RR *(ms)*	Mean RR *(ms)*
2	STD RR *(ms)*	STD RR *(ms)*	STD RR *(ms)*	STD RR *(ms)*
3	RMSSD *(ms)*	RMSSD *(ms)*	RMSSD *(ms)*	RMSSD *(ms)*
4	HRVti	HRVti	HRVti	HRVti
5	TINN *(ms)*	TINN *(ms)*	TINN *(ms)*	TINN *(ms)*
6	VLF power *(log)*	VLF power (ms^2)	VLF power (ms^2)	VLF power *(log)*
7	HF power *(log)*	LF power *(log)*	LF power (ms^2)	LF power *(log)*
8	LF/HF ratio	HF power *(log)*	HF power (ms^2)	HF power *(log)*
9	SD2/SD1	LF power *(%)*	VLF power *(log)*	VLF power *(%)*
10	ApEn	SD1	LF power *(log)*	SD1
11	SampEn	SD2	HF power *(log)*	SD2
12	HR Max - HR Min	SD2/SD1	VLF power *(%)*	DFA a2

Ideally, it is better to select the features inside the ML pipeline. However, running RFECV inside the pipeline could lead to the selection of very distinct sets of features for every model. Therefore, in order to keep the comparison conditions equal in terms of the number and the type of features selected for all devices, the selection of 12 features was performed outside the pipeline by Recursive Feature Elimination (RFE). The decision to choose RFE with 12 features as the feature selection algorithms for all models was made after comparing the classification accuracy using several different algorithms with sets of 12 and 15, as depicted in Figure 8.2b.

8.2.2. Grid Search

Grid search is one of the most efficient ways for testing several hyperparameter settings and finding the model's optimal hyperparameters. However, it is computationally expensive when the number of combinations in the search space is very high. It becomes even more problematic when dealing with multiple estimators, each requiring different hyperparameter optimization. With nested cross-validation, the computation

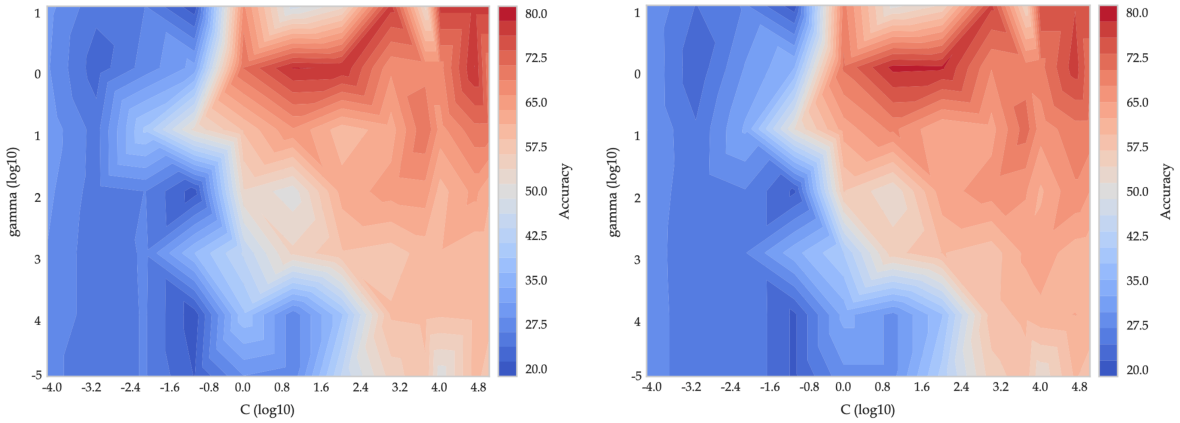


Figure 8.3. Hyperparameter optimization in SVM for (a) Polar H10, (b) Empatica E4.

time increases even further. Most studies use random search for tuning the hyperparameters since it is less expensive. However, we did not want to compromise the quality and reliability of our results to achieve quick results and much higher speeds of operations. Figure 8.3 illustrates the effects of different combinations of hyperparameters on the SVM classifier's output accuracy for two different devices. These two devices are Polar H10, and Empatica E4 depicted in Figures 8.3a, and 8.3b respectively. Despite the overall similarity in both schemes, contour areas are marginally different, and even these slight differences can result in significant changes in the final accuracies. Therefore, by defining device-specific personalized ranges for the parameter grids of each model, we are more likely to achieve the most optimized results for each device. In other words, it is preferable to configure the hyperparameters for each device separately due to the fact that a general model that covers all devices may not show the maximum performance of some of the devices. It should be noted that certain unwritten rules must be taken into account for defining the parameter grid ranges. For instance, using an overly broad range that can significantly increase the chances of wasting time and resources with not much gain in return, and also using specific ranges of values that can cause overfitting must be strictly avoided.

8.2.2.1. Nested Cross-validation. Using the k-fold cross-validation method, we can estimate how well ML models perform when it makes predictions on data not seen during training. Both hyperparameter optimization and model comparison and selection can be achieved using this procedure. However, cross-validation uses the same set of data when tuning the hyperparameters and evaluating model performance. This can result in data leakage and lead to unintended overfitting of the model and overly optimistic biased results [216]. As we have already emphasized the importance of preventing data leakage and ensuring the fairness and reliability of the results, necessary steps had to be taken accordingly. The data leakage problem and the optimistic bias caused by it can be avoided by nesting the hyperparameter optimization procedure beneath the model selection [217]. This procedure is known as nested cross-validation. A nested cross-validation strategy allows for a more robust and generalized assessment of the model performance [218]. It consists of two nested loops, the inner cross-validation loop responsible for hyperparameter optimization, and model selection is nested within the outer CV loop responsible for estimating the generalization error. The inner loop is used for *GridSearchCV* object and the *cross_val_score* object uses the outer loops.

8.3. Classifier Selection

Based on initial experimentation with eight different algorithms, we chose four that showed the most promising results in stratified cross-validation. The first classifier used in this study for stress classification is the Support Vector Machine (SVM) with the radial basis function (RBF) kernel. SVMs are among the most reliable methods in supervised learning algorithms. In the SVM classifier, a point is plotted in the n-dimensional space (n = number of features) for each data item, with each feature's value representing a value of a specific coordinate. As well as being effective in high-dimensional settings, SVMs are versatile since they allow different Kernel functions to be customized for the decision function. SVMs perform best when C and gamma are chosen appropriately. In addition to the utilization of *GridSearchCV* for choosing the best C and gamma values, we carefully examined CV scores and selected the search space values to prevent overfitting and efficiently combine the parameters. The sec-

ond classifier selected to be employed, a random forest classifier, fits many decision tree classifiers on different subsamples of the dataset and combines the averages of the results in order to prevent overfitting and improve prediction accuracy. The random forest output is the class chosen by the majority of trees [219, 220]. As the name suggests, randomness is the prime feature of Random Forests. The concept of randomness was introduced to increase the generalization as a successful attempt to address the problem of decision trees tending to overfit. Randomization is achieved by training each tree solely on a random subspace of samples with a random subset of features pulled from the training set (with replacement).

An additional randomization step results in Extremely Randomized Tree (ExtraTree), which is the third classifier we have employed in this study. Being ensembles of decision trees, both RF and ExtraTree are based on individual decision trees. However, they differ primarily in two ways. In the ExtraTree classifier, the whole learning sample is used to train the tree, and no replacement is done, as opposed to RF, which bootstraps the samples. Moreover, instead of optimum splits, which are commonly done based on the Gini impurity or information gain in RF, in the ExtraTree, a randomized top-down split is used [221].

Gradient boosting is a powerful ML method built as an ensemble of weak learning models. This method relies on the idea of sequentially building models, and those models must try to reduce the errors of the preceding model [222]. Individual decision trees are the weak learners in gradient boosting decision trees. All individual decision trees are connected in series, and each tree attempts to minimize the error of the preceding one. Light Gradient Boosting Machine (LightGBM) is the fourth algorithm used in our study. It is an open-source gradient boosting framework that increases the model's speed and efficiency. It reduces its memory consumption by following a leaf-wise tree growth approach and utilization of two additional novel methods, One-side sampling and exclusive feature bundling [222, 223]. Faster training speed, and lower resource requirements, make LightGBM one of the best choices when frequent retraining or fast evaluation of large datasets are required.

This chapter primarily aims to assess and compare stress detection performance between different wearable devices using supervised ML algorithms. At first, we decided to perform the study only with SVM and RF. However, after observing promising results with RF, we decided to add two additional, more robust decision tree-based methods to the study.

8.3.1. Reproducibility and Hyperparameter Optimization

Considering that this study was conducted to compare the performance of different devices on stress detection accuracy, all comparisons are expected to be fair. As mentioned earlier, a fair comparison requires identical conditions between all models. A key component in fulfilling this fairness was to keep the random_state equal in all models. To accomplish this, we used an identical random_state value globally throughout all operations, from primary data shuffling to splitting the data to training and test sets, as well as the randomness of internal operations of each classifier (e.g., for generation of pseudo-random number for shuffling the data in support vector machine, or for controlling the randomness of the bootstrapping of the samples in use while building the trees in random forest classifier).

In order to achieve the best results, the grid search in our machine learning model pipeline searched for hyperparameters for all seven wearable devices individually. Due to the fact that we utilized four different machine learning algorithms, we obtained a total of twenty-eight different sets of hyperparameters. For example, the hyperparameters obtained by the Random Forest classifier using the data collected with BITalino (r)evolution kit were: 'bootstrap': False, 'criterion': 'gini', 'max_depth': 50, 'max_features': 'auto', 'max_leaf_nodes': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100 , and finally, the type of data scaler selected was MinMax.

8.4. Classification Results

It is common in the literature that many studies only report the test accuracy results. However, there are studies in which decisions must be taken regarding issuing intervention instructions or even a simple notification. When such decision-making is directly related to human health, which is "stress" in our case, it is essential to know the model performance in terms of true/false positives and negative reports as well. For this reason, we have reported several metrics, including accuracy, precision, recall, and F_1 score. Furthermore, results of the cross-validation accuracy on training data are also reported.

Table 8.3 represents the classification results. When we do not have a large dataset, it is feasible to evade splitting the data into train and test and only perform cross-validation on the whole data [224]. However, as already described in Subsection 8.2.2.1, by using the regular cross-validation, the same dataset will likely be used to tune and select a model, which will lead to a biased assessment of the model performance. In order to minimize this bias, model selection should be treated as an integral element of the model fitting procedure, and independent trials should be conducted to avoid selection bias and to exhibit best practices. For this reason, a nested CV is preferred over a non-nested CV to overcome the performance evaluation bias [224]. In the case of using nested cross-validation, applying it to the whole data would be sufficient to report an unbiased estimation of the model performance. Nevertheless, we still retained parts of the data as our holdout test set. This was done to observe the performance of all of our 28 models faced with completely new data that did not exist during the training process and to eliminate any uncertainty regarding the validity of the reported results. These test sets were utterly intact from the onset and had no role neither in model selection and tuning nor in feature selection. The split of our training and testing data in a stratified manner consisted of eighty and twenty percent of the total data, respectively. It should be noted that by the experimental implementations of the nested cross-validation on the whole data, we could achieve an average of three to six percent increased performance in all models. This was due to the fact that in that case, since no

Table 8.3. Classification results using four algorithms for all seven wearables.

		SVM			Random Forest			ExtraTree			LightGBM		
Device	**Session**	**Precision**	**Recall**	**F_1**	**Precision**	**Recall**	**F_1**	**Precision**	**Recall**	**F_1**	**Precisio**	**Recall**	**F_1**
Firstbeat Bodyguard 2	Baseline	80.73	78.72	79.71	83.58	79.43	81.45	86.57	82.27	84.36	82.85	80.50	81.65
	Stress	86.55	84.40	85.46	82.13	84.75	83.42	83.28	88.30	85.71	81.44	84.04	82.72
	Relaxation	73.27	79.48	76.25	79.30	81.11	80.19	81.55	82.08	81.82	77.42	78.18	77.80
	Cycling	94.33	86.93	90.48	90.73	89.54	90.13	92.57	89.54	91.03	93.96	91.50	92.72
	CV Accuracy	79.40% +/- 3.28			80.96% +/- 1.86			83.30% +/- 2.26			80.67% +/- 2.30		
	Test Accuracy	79.69%			82.81%			**86.72%**			82.03%		
Polar H10	Baseline	78.80	79.08	78.94	80.57	80.85	80.71	83.88	81.21	82.52	82.26	77.30	79.71
	Stress	85.56	81.91	83.70	84.23	83.33	83.78	84.25	87.23	85.71	81.23	84.40	82.78
	Relaxation	75.08	80.46	77.67	78.90	79.15	79.02	84.04	84.04	84.04	79.05	81.11	80.06
	Cycling	99.30	92.16	95.59	93.51	94.12	93.81	96.05	95.42	95.74	97.35	96.08	96.71
	CV Accuracy	80.47% +/- 2.06			81.84% +/- 2.23			84.47% +/- 1.89			82.23% +/- 3.33		
	Test Accuracy	80.86%			84.38%			**84.38%**			83.98%		
Zephyr HxM	Baseline	79.41	76.60	77.98	83.21	79.08	81.09	85.82	83.69	84.74	82.80	81.91	82.35
	Stress	86.14	81.56	83.79	84.01	87.59	85.76	86.71	87.94	87.32	85.46	85.46	85.46
	Relaxation	70.99	82.08	76.13	81.43	81.43	81.43	85.25	84.69	84.97	80.07	79.80	79.93
	Cycling	96.92	82.35	89.05	94.19	95.42	94.81	93.04	96.08	94.53	92.36	94.77	93.55
	CV Accuracy	80.27% +/- 2.51			83.50% +/- 2.32			84.47% +/- 1.87			82.13% +/- 0.87		
	Test Accuracy	81.25%			84.38%			**87.89%**			84.38%		
Bitalino	Baseline	74.44	80.16	77.19	81.32	89.88	85.38	86.79	93.12	89.84	82.54	84.21	83.37
	Stress	83.40	80.78	82.07	85.89	83.53	84.69	91.09	88.24	89.64	86.06	84.71	85.38
	Relaxation	78.10	76.98	77.54	87.21	80.94	83.96	88.52	85.97	87.23	79.93	80.22	80.07
	Cycling	93.94	89.21	91.51	92.14	92.81	92.47	94.89	93.53	94.20	91.97	90.65	91.30
	CV Accuracy	79.00% +/- 2.08			84.88% +/- 2.94			87.16% +/- 2.40			84.01% +/- 2.13		
	Test Accuracy	80.43%			84.78%			**88.26%**			85.65%		
Empatica E4	Baseline	78.97	75.89	77.40	82.35	79.43	80.87	84.36	82.27	83.30	78.82	80.50	79.65
	Stress	81.25	78.37	79.78	82.44	81.56	82.00	79.66	83.33	81.46	82.48	80.14	81.29
	Relaxation	68.34	75.24	71.63	76.71	80.46	78.54	79.80	79.80	79.80	77.24	78.50	77.87
	Cycling	92.31	86.27	89.19	94.70	93.46	94.08	95.24	91.50	93.33	93.33	91.50	92.41
	CV Accuracy	75.78% +/- 0.82			80.28% +/- 3.45			81.35% +/- 3.11			79.88% +/- 2.67		
	Test Accuracy	81.25%			**83.98%**			82.42%			81.25%		
Samsung Gear S2	Baseline	77.06	76.24	76.65	82.56	82.27	82.42	85.47	87.59	86.51	81.36	80.50	80.93
	Stress	78.76	72.34	75.42	83.15	82.27	82.71	86.13	83.69	84.89	78.32	79.43	78.87
	Relaxation	69.54	78.83	73.89	80.00	84.69	82.28	82.54	84.69	83.60	79.10	80.13	79.61
	Cycling	87.68	79.08	83.16	94.24	85.62	89.73	91.10	86.93	88.96	89.86	86.93	88.37
	CV Accuracy	75.30% +/- 2.08			79.40% +/- 1.93			83.01% +/- 0.92			78.12% +/- 2.80		
	Test Accuracy	80.08%			83.59%			**83.98%**			79.30%		
CorSense	Baseline	78.86	83.63	81.17	79.74	86.83	83.13	80.72	87.90	84.16	81.19	87.54	84.25
	Stress	81.82	83.27	82.54	81.72	84.34	83.01	83.28	86.83	85.02	85.36	85.05	85.20
	Relaxation	81.06	69.48	74.83	80.83	62.99	70.80	85.47	64.94	73.80	79.70	68.83	73.87
	CV Accuracy	79.74% +/- 3.63			79.05% +/- 3.25			81.98% +/- 3.47			80.17% +/- 3.34		
	Test Accuracy	**84.44%**			82.22%			80.56%			80.56%		

data was reserved for the test, consequently, more data was available for training. Since this study is primarily devoted to the comparison of the stress detection performance across multiple wearables, we will not be focusing on the models and comparing the algorithms in great detail. However, a cursory glance at Table 8.3 suggests that,

overall, ExtraTree shows promising results and is proving to be more effective than the other algorithms. On the other hand, SVM appears to be the least effective of the four classifiers in this study. Further visual inspection of the table indicates that the Random Forest and LightGBM algorithms appear to have performed almost equally well as the ExtraTree algorithm. As described in Section 8.3, since the ExtraTree can be considered as an enhanced version of Random Forest, in order to avoid increasing the number of detailed comparisons, we will continue this section by closer inspection of device performances with two algorithms, namely LightGBM, and ExtraTree classifiers. As expected, wearables equipped with ECG sensors that can record raw data with higher quality [19] maintained their superiority in stress classification applications as well. In the ExtraTree classifier, the average test accuracy of ECG and PPG wearables is 86.81% and 82.32%, respectively. Similarly, for LightGBM, it is 84.01% and 80.37%. These results indicate that ECG wearables performed 5.45% and 4.52% better than PPG wearables with ExtraTree and LightGBM models. In order to examine the results of the different classes in more detail, the Precision, Recall, and F_1 Scores for each class are also accessible using this table. These results are obtained by averaging the values of these metrics obtained from the outer folds of the nested cross-validation and are a valid criterion for presenting the performance of models on a large portion of the data. The class-wise comparison of these metrics in the top two algorithms shows that almost all devices score above 80% on all three metrics. We observe excellent results in the physical stress class (Cycling). It proves that the magnitudes of changes in the HRV features while performing rigorous physical activity are so intense that nearly all metrics for this class achieve scores above 90%. It was anticipated that all models would face difficulty choosing between recovery and baseline classes because of the remarkable similarity. However, despite the fact that compared to other classes, we can see slightly lower performances in these two classes in all models, our two top-performing models classify these two classes with excellent scores.

The results in Table 8.3 provide a brief overview of the overall differences between our seven devices. Moreover, the normalized confusion matrix for all models is depicted in Figure 8.4. However, we need to take further steps to make a valid judg-

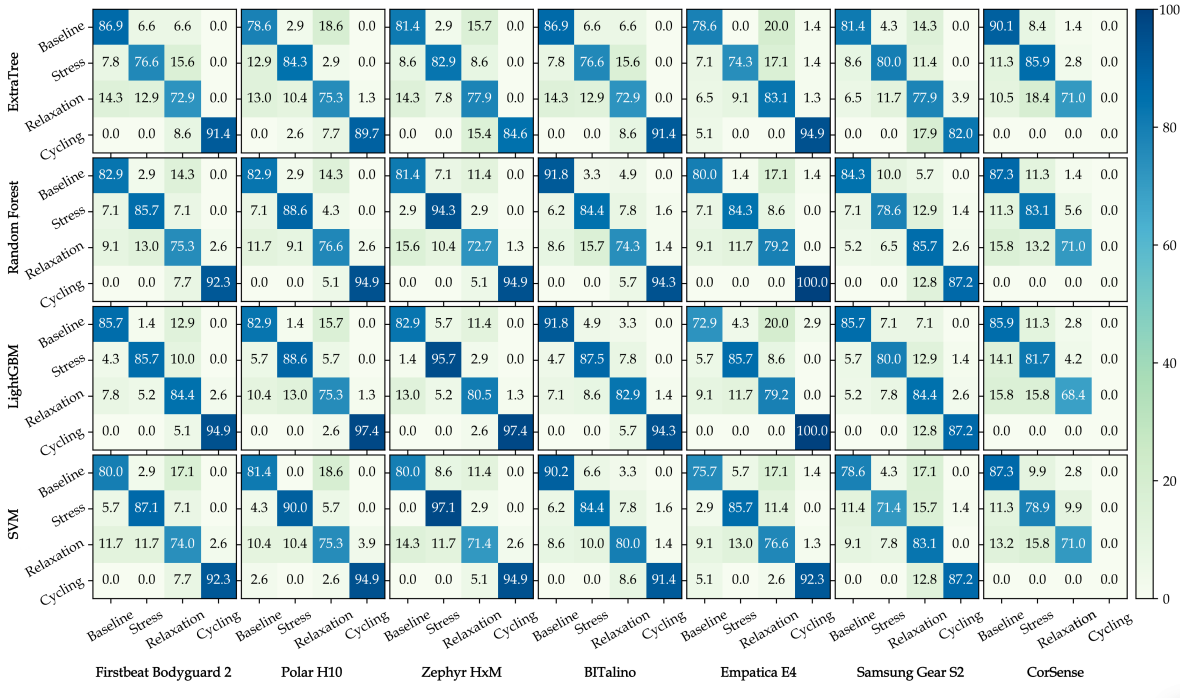


Figure 8.4. Normalized confusion matrices using four classifiers for all devices.

ment and a final statement. Since comparing all of the 28 models with at least five metrics in each can cause unnecessary confusion and create a new perplexing problem, we make a statistical comparison of all the metrics of the two top algorithms cumulatively. Following Shapiro-Wilk and Levene's tests to examine the normality of distributions and equality of variances in our results, since the assumptions of normality were not met, we decided to utilize the Kruskal-Wallis test [186, 187]. As seen in Figure 8.5, in the LightGBM models, Kruskal-Wallis showed a significant main effect ($p = 0.009$). Hence we continued with posthoc analysis. For this, pairwise comparisons were performed with Dunn's test, and p-value adjustment following multiple pairwise comparisons was carried out using Holm's method [225]. For the results obtained with LightGBM for all classes, posthoc analysis showed significant differences only between BITalino (r)evolution and Samsung Gear S2 wearables ($p = 0.028$) (See Figure 8.5a). A similar result was repeated in the stress class ($p = 0.024$). Simply put, using LightGBM, the only statistically significant difference exists between the ECG device with the best results (BITalino (r)evolution) and the PPG wearable with the lowest results

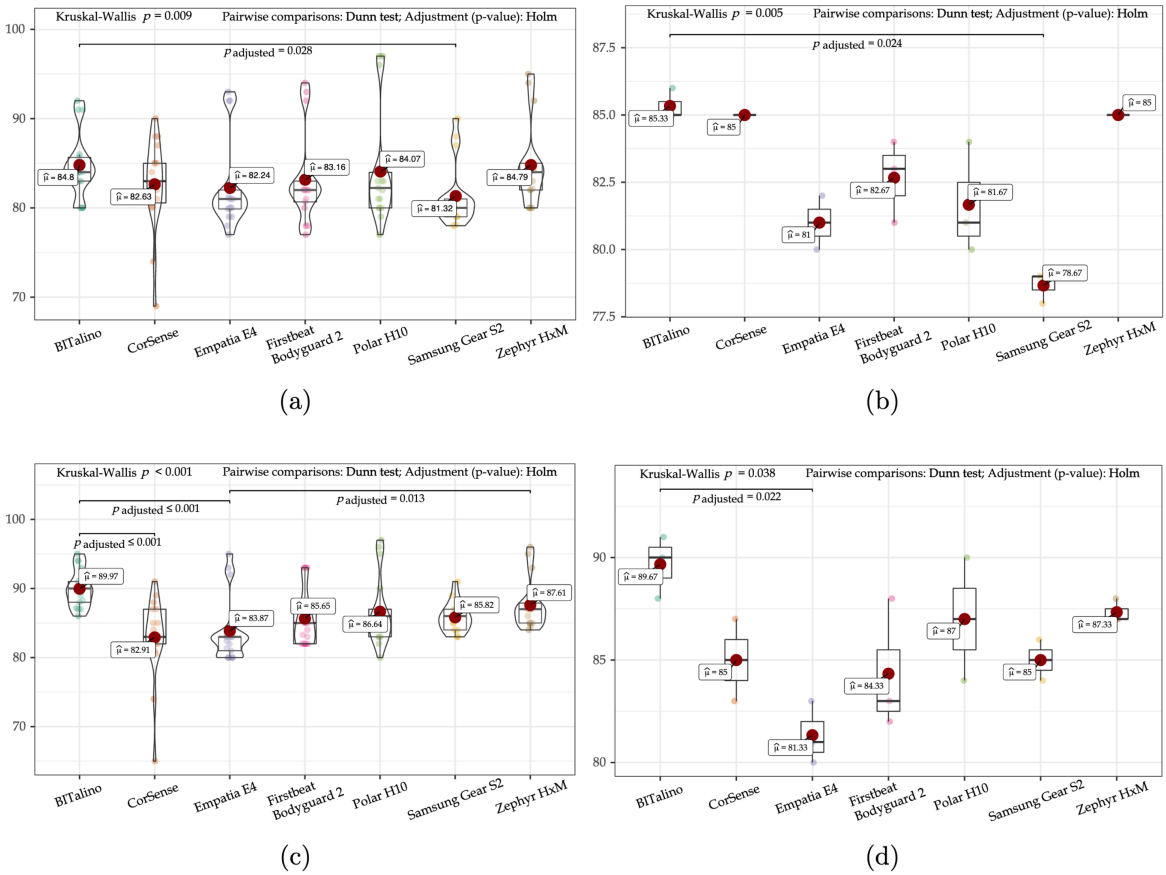


Figure 8.5. Kruskal-Wallis comparison followed by Dunn's test for all devices in (a) All sessions with LightGBM classifier, (b) Stress session with LightGBM classifier, (c) All sessions with ExtraTree classifier, and (d) Stress session with ExtraTree classifier.

(Samsung Gear S2), and there exists no statistically significant difference between other devices (See Figure 8.5b). Repeating the same statistical approach to analyze the results obtained from the ExtraTree classifier results in more conservative results (See Figure 8.5c). Here, in the all-classes comparison, there are significant differences between BITalino (r)evolution and two other PPG devices (Empatica E4 and CorSense, and between Zephyr HxM and Empatica E4 as well. In the stress class, similar to the LightGBM, there is only a statistically significant difference between the two ECG and PPG devices (BITalino (r)evolution and the Empatica E4) (See Figure 8.5d).

8.4.1. Effects of Multimodality

Wearable devices in this study showed great potential for producing reliable stress detection and classification results, especially those equipped with ECG sensors. However, up until this point, all comparisons in this study were made solely based on HRV data calculated from cardiac signals. Considering the fact that HRV is a valid criterion for detecting psychophysiological changes, a critical question to address is whether collecting the data solely from a single modality (cardiac signals in our case) would be sufficient for stress detection. Since user expectations from different applications may differ, an appropriate answer can be that it would be best for the end-user to decide on this issue, keeping both the pros and cons of utilizing multimodality in mind. For instance, by employing multimodality, depending on the availability of multiple sensors on a device, we will be limited to using a particular type of device or even a combination of two or more devices simultaneously. In addition to bringing unobtrusiveness, this will lead to extra computational load and higher energy consumption. In this section, we will investigate the effect of multimodality by comparing the classification results with and without multimodality.

Among all devices employed in this study, only the Empatica E4 is equipped with an EDA sensor capable of capturing EDA biosignals at a rate of 4 Hz. To bring another modality into the study, we first performed the necessary preprocessing and feature extraction on this EDA data (see section 4.7.2). As seen in Figure 8.6, the Tonic and Phasic components of the EDA signals were extracted using NeuroKit2, and a closer look at the EDA data from five subjects is demonstrated at the bottom of the exact figure.

Based on the results reported in Table 8.4 and displayed in Figure 8.7, we achieved a significant improvement in classification results with the addition of a single additional modality (EDA). The Empatica E4 (with EDA) shows the highest accuracy of all devices. This record-breaking increase includes all models, with a staggering 90.62% accuracy with the ExtraTree algorithm. These results indicate that a wearable with

Table 8.4. Classification results for the Empatica E4 with and without EDA.

		SVM			Random Forest			ExtraTree			LightGBM		
Device		**Precision**	**Recall**	**F_1**	**Precision**	**Recall**	**F_1**	**Precision**	**Recall**	**F_1**	**Precisio**	**Recall**	**F_1**
Empatica E4	Baseline	78.97	75.89	77.40	82.35	79.43	80.87	84.36	82.27	83.30	78.82	80.50	79.65
	Relaxation	68.34	75.24	71.63	76.71	80.46	78.54	79.80	79.80	79.80	77.24	78.50	77.87
	Stress	81.25	78.37	79.78	82.44	81.56	82.00	79.66	83.33	81.46	82.48	80.14	81.29
	Cycling	92.31	86.27	89.19	94.70	93.46	94.08	95.24	91.50	93.33	93.33	91.50	92.41
	CV Accuracy	75.78% +/- 0.82			80.28% +/- 3.45			81.35% +/- 3.11			79.88% +/- 2.67		
	Test Accuracy	81.25%			**83.98%**			82.42%			81.25%		
Empatica E4 HRV + EDA	Baseline	76.35	80.14	78.20	83.99	83.69	83.84	85.05	84.75	84.90	85.92	86.52	86.22
	Relaxation	74.92	73.94	74.43	80.19	81.76	80.97	82.79	83.06	82.93	81.41	82.74	82.07
	Stress	79.20	76.95	78.06	86.28	84.75	85.51	86.11	87.94	87.02	86.45	83.69	85.05
	Cycling	91.39	90.20	90.79	93.46	93.46	93.46	95.24	91.50	93.33	90.97	92.16	91.56
	CV Accuracy	77.15% +/- 1.90			83.40% +/- 3.46			84.67% +/- 3.86			83.40% +/- 4.23		
	Test Accuracy	82.81%			89.84%			**90.62%**			88.28%		

a single type of sensor (ECG), regardless of how well it records high-quality HRV data and leads to high classification accuracy, is still not a silver bullet. Moreover, with respect to the fact that HRV is a robust criterion for detecting stress, bringing a second type of sensor to the table, and exploiting multimodality can lead to much better results.

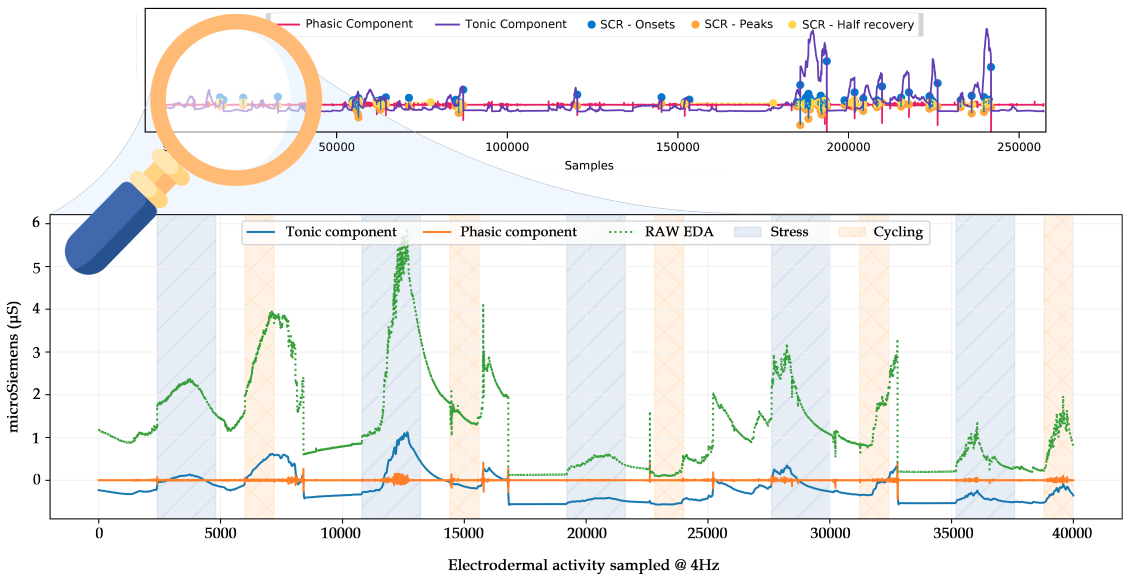


Figure 8.6. Sample of electrodermal activity for five participants.

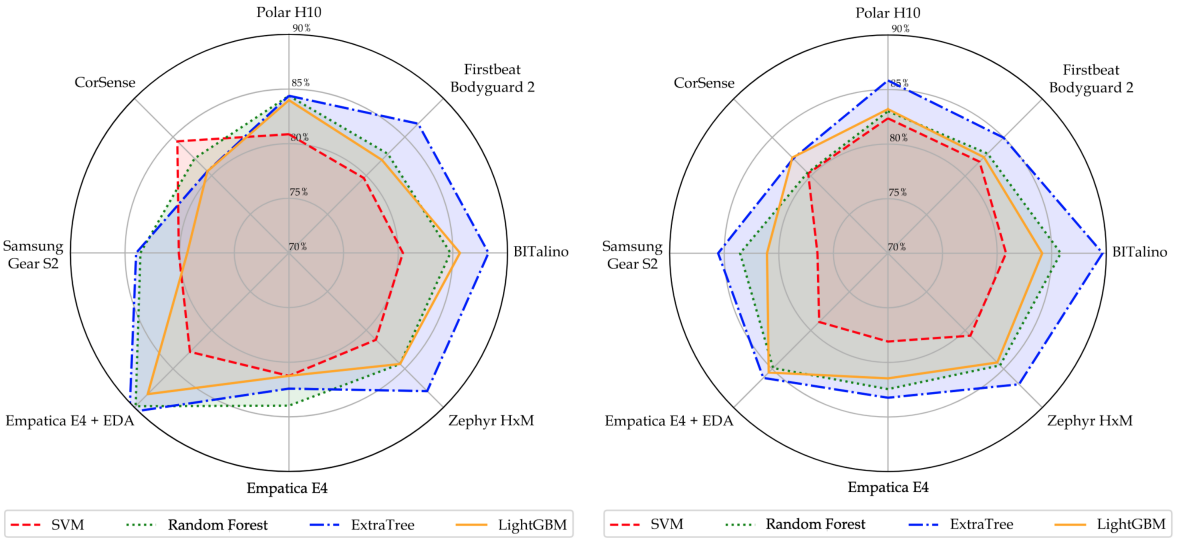


Figure 8.7. Comparison of (a) Test accuracy and (b) F_1-Score performance using four classification algorithms on data from seven devices.

8.5. Model Explainability

In many applications, understanding why a model reaches a particular prediction is just as essential as its accuracy. Nowadays, achieving high accuracies with very complex models is often possible. However, the model behavior and identifying the factors involved in the outcome becomes very hard to interpret [226] (see Figure 8.8). The ever-growing application of black-box machine learning models leads to the crucial need for justifying and interpreting their decisions. This challenge is a significant barrier to ML adoption in critical applications, such as healthcare. Even though ML models have made it considerably easier to predict the feature health conditions of an individual, they still fall short in interpretability [227]. Identifying and interpreting which features contribute most to a particular prediction in different models can be very useful, mainly if such analysis can be applied to specific classes and individuals. Therefore, model interpretability can become crucial in ML problems related to early detection and intervention in human health [228]. An essential aspect of investigating the models' explainability in the context of this study is to determine whether the

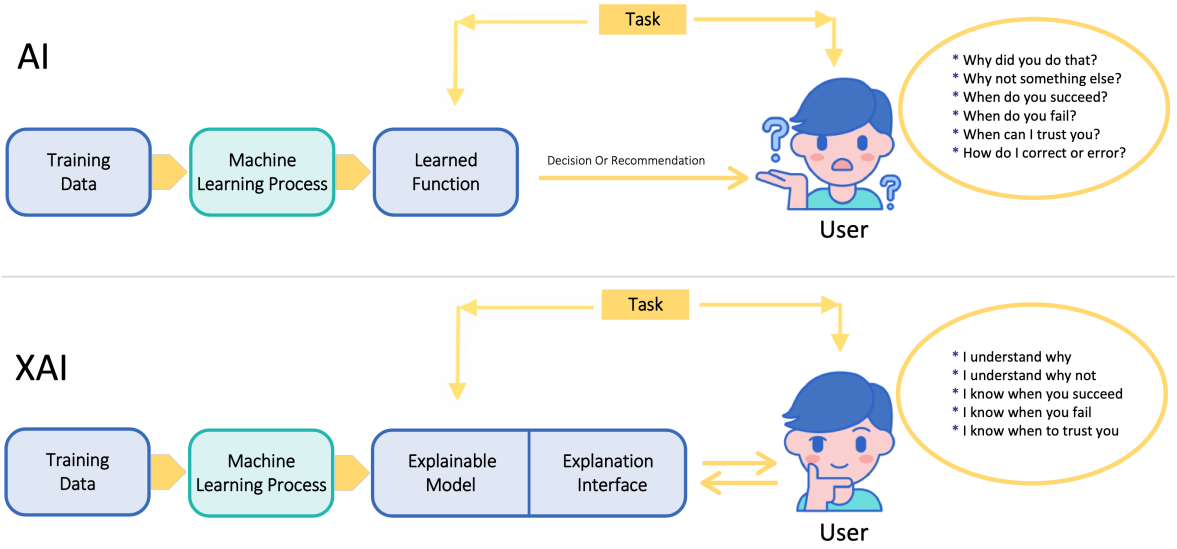
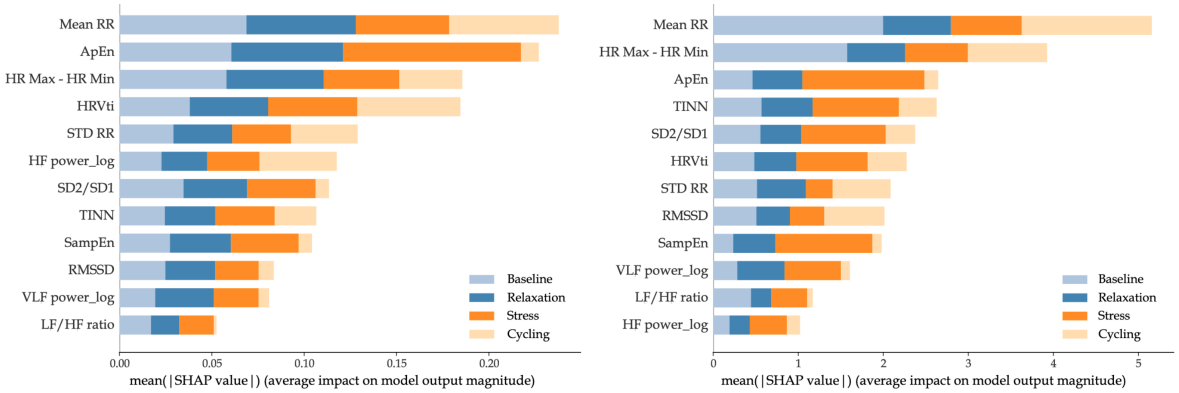


Figure 8.8. Explainibility of AI, helping clinician, psychologists, and-users better understand the model.

same features from devices with different sensing technologies (ECG vs. PPG) have similar effects on model outputs and classification results. Lime, Dalex, and SHAP are examples of analytical tools in the area of Explainable Artificial Intelligence (XAI) [226], [229, 230]. These tools have evolved for model interpretation and demystifying black-box models over the last few years and are becoming more popular each day.

8.5.1. SHapley Additive exPlanations (SHAP)

In this study, we utilize SHapley Additive exPlanations (SHAP) for explaining our models. SHAP is an open-source game-theoretic approach for explaining the results of ML models. It can probably be considered state-of-the-art in XAI. Shapley value is a term used in game theory. It is a solution concept named in honor of Nobel Prize-winning economist Lloyd Shapley. Derived from the Shapley values of the original model's conditional expectation function, SHAP values are unified measures of feature importance [226]. Applying SHAP analysis to our data and model outputs results in the production of matrices containing the SHAP values. These matrices are in the same dimension as the original data matrix. To provide a better comprehension of



(a) (b)

Figure 8.9. Feature influences with SHAP on all classes, with (a) ExtraTree, and (b) LightGBM models using the Firstbeat Bodyguard 2 wearable device.

the subject for the readers who may not be familiar with the game theory and the above concepts, it would be beneficial to provide a brief example for the whole concept and extend it to its application in this study [231]. The game theory requires at least two elements: a game and its players. Assuming we have a classification model, the "game" would be responsible for producing the model's results. In this example, the "players" have the role of features in our model. Shapley quantifies each player's contribution to the game, while SHAP quantifies each feature's contribution to the prediction made by the model. For instance, in our case, SHAP values can show the effects of the RMSSD feature on each class, and this interpretation can be performed either globally or locally. For the global interpretation, SHAP can show how much each feature impacts the prediction of each class, either negatively or positively. Unlike the traditional feature importance plot, SHAP can generate plots that can demonstrate each feature's positive or negative impacts on the target. In the local interpretation, each observation receives its own respective set of SHAP values [232]. By this means, the interpretation of individual subjects becomes possible. This is a significant increase in transparency compared to the conventional feature importance algorithms that only display the results of the whole population. Using stacked bar plots, Figure 8.9 shows the mean of SHAP values for all features. This is equivalent to the average impact of each feature on the output of two top-performing models, ExtraTree, and LightGBM,

depicted in Figures 8.9a, and b, respectively. Data for these two models came from Firstbeat Bodyguard 2. As seen in both plots, the time-domain feature, Mean RR, shows the highest impact on the overall output of the model in both models. While the nonlinear feature ApEn is in second place with the ExtraTree classifier, it is in third place using LightGBM. Positions of the second and third-ranked features in the two models are in the opposite order. Although we can see changes in several steps in the ranking order of some of the features, from the top of the list to its bottom, there is a high overall similarity between the importance of the features in both models. From a further extensive and class-wise perspective, we can interpret these plots as follows. While in the ExtraTree, Mean RR influences the prediction of each class in almost the same magnitude, in LightGBM, this influence is doubled for the Baseline and Cycling. ApEn has the most significant influence on predicting Stress in both models, and the frequency-domain feature LF/HF has little to no impact on predicting the physical stress (Cycling) session.

As a final example for Figure 8.9, the effect of the SampEn nonlinear feature on the estimation of the Stress class is almost twice the sum of its influences on the other three classes. In summary, these plots allow us to gain an understanding of what our machine learning model has learned from the features. These analyses demonstrate that two different models behave very similarly on the same device and that the identical features in two different models have more or less the same effects on the model output. Nonetheless, there are also some differences in the order of importance of the features in the overall output of the two models and the extent to which they influence these two models in choosing a particular class as an output. This shows that regardless of the type of device used, an in-depth interpretation of the stress level measurement and the importance of the features involved can be strongly influenced by the type of the employed model.

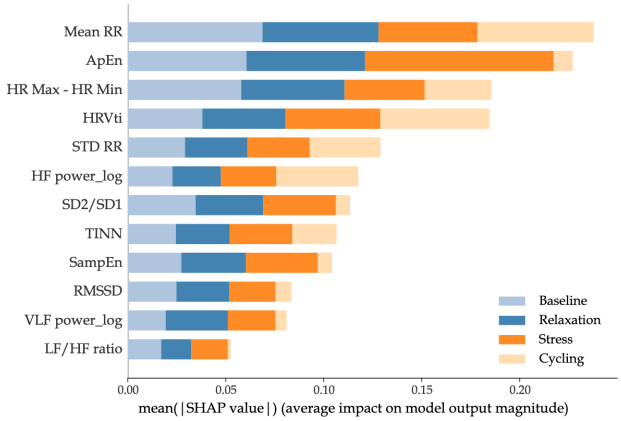
It was theoretically possible to select the best features through SHAP by embedding it in the machine learning pipelines. We could achieve even better classification results by adopting such an approach, but we avoided doing this for two reasons.

Firstly, in that case, we would obtain sets of features that may be different to certain degrees for each classification algorithm and different wearables. Since we arranged all the comparison conditions to be equal to achieve a fair comparison, the set of features used should have been the same as well. Secondly, due to the high computational cost of SHAP, if it was embedded inside our pipelines that were equipped with nested-CV and grid search, which both are very time-consuming on their own, achieving the result in a reasonable time would become beyond the computing power of our equipment.

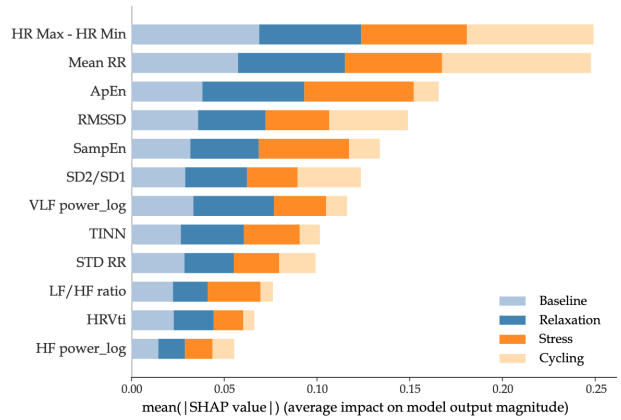
In Figure 8.10, we present a different analogy, comparing two devices with a single classifier. In this comparison, we have two devices of the type ECG and PPG, in Figures 8.10a, and b, respectively. In Figure 8.10c, we present the effect of multimodality on the model's output. There are differences in the order of feature impact rankings in all three plots. The amount of differences seen in 8.10a are naturally greater due to the presence of EDA features. However, a similar pattern can be seen both in the order of the features and in the influence of individual features on the model's output. For example, ApEn has the greatest impact on stress class prediction in all three models, and frequency-domain features are in the last of the rankings. This shows that even similar models behave differently with devices of different types (ECG, PPG, and EDA), and feature importances also show higher differences.

In order to examine the effects of features on each class more precisely, it is necessary to zoom in to a more detailed view. Figure 8.11 shows horizontal scatter plots for each feature with different color gradients. Feature importances and feature effects are aggregated in these two class-wise summary plots. Each point represents a Shapley value for a feature and an observation on these scatter plots. While features are positioned on the y-axis, Shapley values of their instances are positioned on the x-axis. For better visualization, overlapping points are jittered.

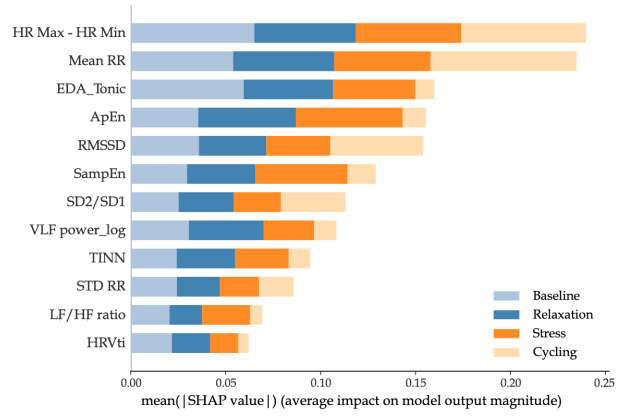
The Intensity and gradient of the colors for each instance indicate the feature values from low (blue) to high (pink), as shown in the color bar on the left side of the plots. We already examined the behavior of ApEn in Figure 8.10. ApEn had the



(a)

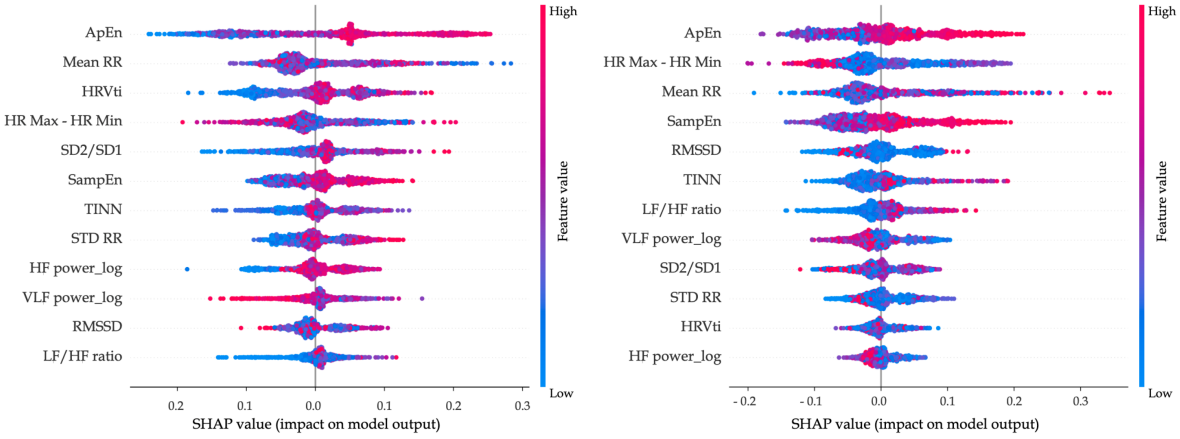


(b)



(c)

Figure 8.10. Feature influences with SHAP on all classes, with ExtraTree classifier using the (a) ECG (Firstbeat Bodyguard 2), (b) PPG (Empatica E4), and (c) PPG + EDA (Empatica E4 + EDA) data.



(a) (b)

Figure 8.11. Feature influences with SHAP for the Stress class, with ExtraTree classifier using (a) Firstbeat Bodyguard 2, and (b) Empatica E4 data.

most impact on predicting the stress class in both devices. We can now investigate the same behavior in a more detailed and class-wise perspective, using Figures 8.11a, and b for Firstbeat Bodyguard 2, and Empatica E4 wearables, respectively. Upon close investigation of the plots in Figure 8.11 we can realize that with higher (more pink) values of ApEn, the model is more likely to classify the class as stress, whereas with lower (more blue) values, it is less likely to do so. In other words, a high (more pink dots) level of ApEn has a high and positive (more towards the right dots) effect on the class, being predicted as stress. In a similar fashion, we can say that (HR Max - HR Min) is negatively correlated with the class being predicted as stress.

In another example of using SHAP to gain a deeper understanding of the results from various models, we examined the effects of different data scaling methods on the outputs obtained from different models. As seen in Figure 8.12, The Random Forest classifier applied to the same sets of data scaled with two distinct types of scalers produces almost identical results in all devices (See Figures 8.12a, and b). There are no visible differences in the classification results, features' importances, or their impact on the classification result. Although not shown in this figure, the ExtraTree classifier is no different and follows the same behavior as well. However, as seen in

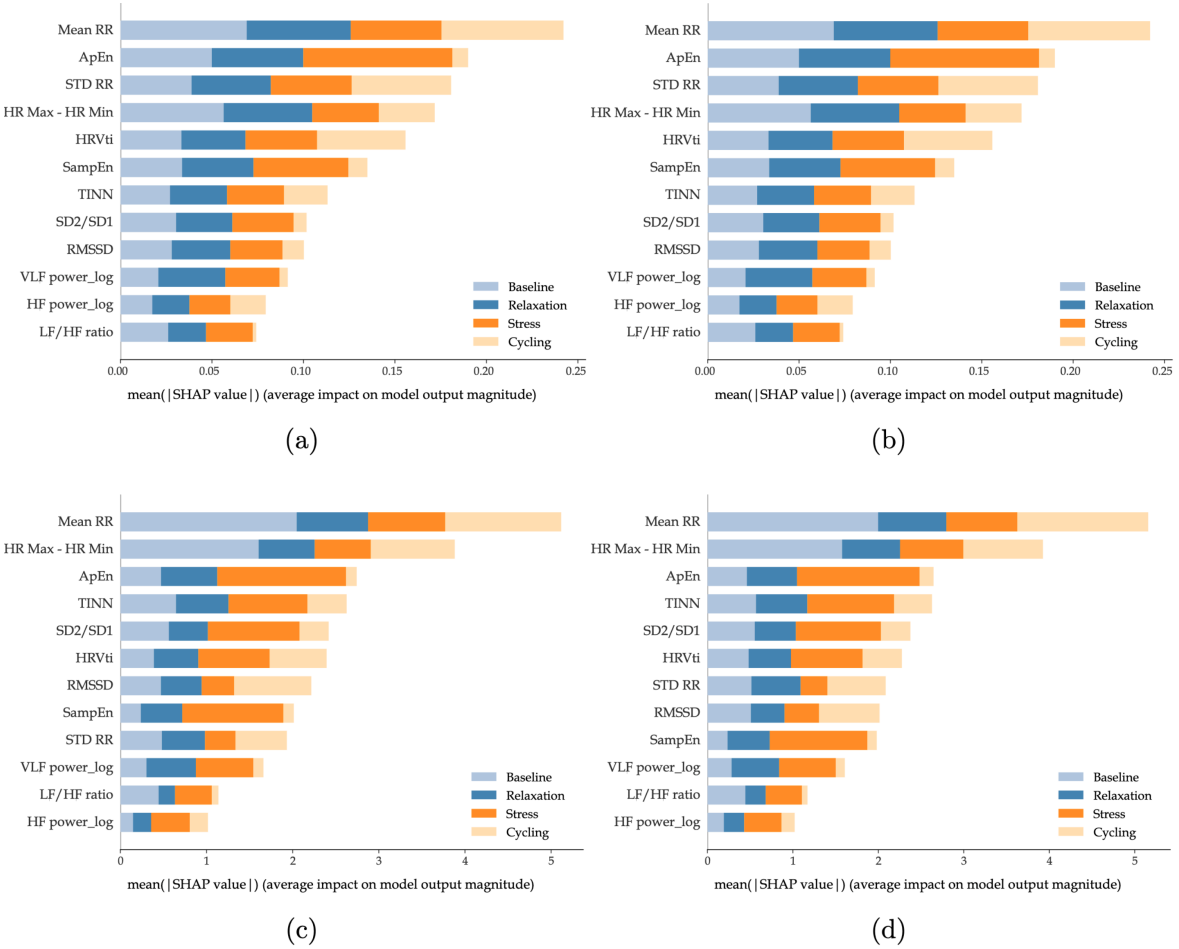


Figure 8.12. Effects of using different types of scaling in model output with (a) Random Forest - Robust scaling, (b) Random Forest - MinMax scaling, (c) LightGBM - Robust scaling, and (d) LightGBM - MinMax scaling.

Figures 8.12c, and d, the LightGBM classifier shows different results for the data scaled with different types of scalers. This is due to the fact that the MinMaxScaler is exceptionally sensitive to the presence of outliers. However, in the RobustScaler procedure, scaling and centering calculations are based on percentiles, and as a result, outliers of a considerable magnitude do not affect the outcome much. Since boosting methods make trees fix the errors made by their predecessors and build each tree on the residuals of previous trees, outliers will have a much larger residual than non-outliers, making LightGBM and other boosting methods, in general, more sensitive to outliers. This shows that such behavior may be caused by combining MinMaxScaler

with a boosting algorithm. In this case, it would be sufficient to minimize the effects of outliers as much as possible, for example, by using a scaling method that is robust to outliers. This shows that some models may also be sensitive to certain preprocessing steps. In such a case, even if accurate classification results were obtained, detailed analysis and study of the effects of different features based on this model will not be very reliable. Prior to implementing ML models, it is necessary to be aware of their possible weaknesses, shortcomings, and compatibilities to avoid any factors that may lead to undesirable predictions by a particular algorithm. All these details have been carefully taken into account in this study, and appropriate measures have been taken to overcome all potential challenges.

In line with the primary objective of this study to compare the performance of wearable devices in stress measurement and to achieve a more robust conclusion, we employed SHAP for model explainability. By doing so, we were able to gauge the performance of the devices more accurately and make much more fair decisions when choosing between them. Furthermore, the purpose of using SHAP for the explainability of our models was to analyze our results and demonstrate the hidden potentials XAI can offer to studies related to affective computing. The highly functional and unique capabilities of SHAP in examining the factors involved in model decision-making and the comprehensiveness of SHAP values as being unified measures of feature importance can provide new opportunities for researchers. Using SHAP, researchers can scrutinize the factors involved in the occurrence, increase, or decrease of mental stress in the general study population or even in a particular individual.

8.6. Summary and Final Thoughts

In this chapter, a total of seven wearable sensors were selected for stress detection in four classes, namely Baseline, Stress, Relaxation, and Cycling. With four traditional machine learning models, precise tunings were carried out to ensure unbiased results. Results showed that statistically significant differences exist between some of the devices in the classification performance. It is a fact that a statistically

significant difference may prove useful for researchers who want to achieve the highest performance in stress level measurement applications and wish to gain insight into the importance and influences of different features. Nonetheless, as far as the end-user is concerned, all devices deliver very similar results, and there is not much difference in the stress measurement performance for the end-user. Consequently, they are all acceptable for daily use. As shown in the previous study [19], the end-users' ultimate decision may be influenced more by wearability and unobtrusiveness than by seemingly minor differences in performance.

As part of this study, we also used SHAP to make our machine learning models explainable. Using SHAP, we showed that there could be differences between models in the way they prefer one class over another. This was because there were differences in the amount of influences that features had on model output in different models. In some respects, this proves that the effects of HRV features on stress reported in similar studies must be taken with a grain of salt. In addition, we also examined the effect of different preprocessing methods on the output of some models.

Our results showed that ECG wearables demonstrate slightly better performance in all our sessions. Nonetheless, since the ultimate goal of this article is to study the comparison of different devices in the context of stress detection; thus, readers might expect specific devices to be announced as the winner of this analysis. However, we must be cautious in announcing the study's findings in order to avoid creating a biased opinion that could lead to a far-fetched theory. It would be possible to render a final verdict on the superiority of a particular device if, under all conditions, very consistent and similar results were obtained. However, this was not the case. We found that the choice of classifier, and even differences in the data normalization and scaling methods, can influence the model's outcome and change the feature's importance.

Furthermore, we found that multimodality improves stress detection performance in a very significant way. While this is true, it still remains difficult to say definitively whether using the best performer single-sensor ECG device or a multi-sensor device

like the Empatica E4 with PPG and EDA sensors is more effective. Last but not least, we observed that all of the devices used in this study showed relatively high and nearly similar performance in the stress detection application. As a result, the final decision for choosing a particular wearable device over another can be based on the inclusion of additional factors such as personal preferences, expectations, and the pros and cons of each device.

9. CONCLUSION

In the first chapters of this thesis, we explained in detail the basics such as terms related to emotion and affect recognition and their regulation. In addition, we explained concepts such as the types and properties of sensors and actuators used in emotion detection, biofeedback, and reregulation. In the following, in order to further reveal the objectives of this thesis and help the reader gain a better understanding of the subject, we listed some of the similar studies in the literature. Our methodology, which included the types of devices we used, how we collected data, and how we processed data, was explained in Chapter 4 in detail.

An unobtrusive and smart mechanism was implemented in Chapter 5 to detect high stress levels and suggest appropriate relaxation methods (e.g., traditional or mobile) when users face high levels of stress in their daily lives. Also, we developed a system to suggest relaxation methods based on the user's physical activity and environment. While the majority of studies in the literature only detect individuals' stress levels and have no mechanism to recommend regulation techniques as an intervention, our proposed mechanism measures participants' stress levels and helps them regulate their stress levels using a series of popular practices such as yoga, mindfulness, and mobile-based mindfulness apps. The results of this study indicate that yoga and traditional mindfulness perform better than application-based mindfulness methods that are based on mobile phones.

Chapter 6 was the result of a study in cooperation with HCI researchers, in which users' priorities and preferences were put first by letting them develop patterns for regulating emotions based on their preferences. By involving the subjects in the design and personaliziation of haptic patterns, Chapter 6 examined haptic modalities for emotion regulation. During the stress induction tasks, where stress and heart rate variability were evaluated, it reports on the perceived characteristics of these haptic patterns as well as their influence on emotion regulation. The study indicates that

subjective and objective measures of stress were decreased under haptic patterns than without them. Low-frequency vibrations were the most effective at regulating stress. Due to the constraints of the two devices, the ability to personalize vibrotactile and thermal patterns was limited to a certain extent. The experiment, however, was conducted in a laboratory setting and should be evaluated in a real-life stressful scenario. These findings offer new design possibilities for affective well-being regulation technologies, including designs for thermal and vibrotactile biofeedback with personalized and dynamically adjusting patterns.

Having introduced several factors contributing to a reduction in signal quality, we conducted a quantitative and qualitative analysis of different wearables in Chapter 7. We showed that if researchers and end-users intend to use wearable devices for daily use or research, they need to be aware of the strengths and limitations of each wearable and consider the trade-offs between their preferences and the actual capabilities of each device. Additionally, we showed that environmental noises could have a very destructive effect on signals and make the data recorded by some devices useless. However, researchers can dispel the myth that PPG wearables are unreliable for recording and analyzing HRV data by choosing appropriate noise reduction techniques.

In Chapter 8, we discussed critical issues regarding ethical concerns from the data point of view, the confidentiality of human health data, and the great importance of the accuracy and fairness of the diagnoses issued by machine learning models in applications related to mental well-being. We also discussed the methods to avoid biased conclusions and showed how even minor things such as choosing the type of data normalization algorithm and choosing the model's hyperparameters can affect the model's output dramatically.

Furthermore, we showed how we can make our black box ML model understandable for the end users, clinicians, and psychologists so that they can easily understand the cause and effect relationships of different features and classes by utilizing a state-of-the-art explainable AI method.

Finally, and following the various studies that we have done in the scope of this thesis, we believe that by properly aggregating a combination of ideas and mechanisms implemented in this thesis, it would be possible to implement a fully functional and efficient stress detection mechanism that is unobtrusive in terms of hardware and reliable and always available in terms of software. The final product would definitely be a combination of things such as proper knowledge regarding the selection of the type of sensor and wearable, selection of biosignals and features with the greatest impact on ML model output, optimal utilization of the system-labeled data extracted from the context information and the labels set by users, implementation of an accurate and precise stress detection model with the ability to adapt to the environment and the user, implementing effective feedback and intervention system to suggest both traditional practices like yoga, and newer technologies such as haptics, integrating the last two, and finally encouraging the users and public to use such devices. It must be noted that encouraging people to use such a system will ultimately lead to the main goal of the research and investment in this area. We believe that by motivating the users to participate in the design of different stages of the detection and intervention systems, the final product will become more user-friendly, more wearable, and eventually be used more. Moreover, to close the gap between psychologists, computer scientists, and HCI experts and to help them increase their mutual understanding of each other and their interdisciplinary collaborative projects in affective computing, a common and more understandable language should be put into use. We believe that the use of explainable artificial intelligence by us, that is, computer science researchers, will be a big step towards achieving this goal.

One of the limitations regarding the studies conducted in this thesis, which is also prevalent in similar works, is that the number of subjects and devices available for data collection is limited due to financial, human, and time constraints in academic research groups. This issue can lead to specific problems in the future. For instance, the amount of collected data is not sufficient for applying deep learning algorithms, and in case of data loss from any user or their withdrawal from participation, it will be almost impossible to compensate for the lost data. Another limitation of our study in

Sections 7, and 8 is that it was impossible to collect data in real life due to the variety of devices. No matter how unobtrusive our wearables are, users could not carry out their daily life routine with seven devices simultaneously connected to their bodies.

The main purpose of some of the studies in this thesis, which were conducted in a laboratory environment, was to compare different wearables. Therefore, in order to achieve a fair comparison, a level playing field must be provided for all users and devices. Since it was almost impossible to achieve such conditions in uncontrolled environments outside the laboratory, we had to choose the laboratory to conduct the experiments. The other models used in this thesis (apart from the models in Chapter 8) have all passed their tests in everyday life. In order to check the daily life effectiveness of the laboratory models mentioned in Chapter 8 of this thesis, they need to be tested in daily life, which can be an interesting topic for a future study.

REFERENCES

1. Pickering, T. G., "Mental Stress as a Causal Factor in the Development of Hypertension and Cardiovascular Disease", *Current Hypertension Reports*, Vol. 3, No. 3, pp. 249–254, 2001.

2. Schneiderman, N., G. Ironson and S. D. Siegel, "Stress and Health: Psychological, Behavioral, and Biological Determinants", *Annual Review of Clinical Psychology*, Vol. 1, p. 607, 2005.

3. Picard, R. W., *Affective Computing*, MIT press, London, 2000.

4. Brown, B. B., *Stress and the Art of Biofeedback*, Harper & Row, New York, 1977.

5. Frank, D. L., L. Khorshid, J. F. Kiffer, C. S. Moravec and M. G. McKee, "Biofeedback in Medicine: Who, When, Why and How?", *Mental Health in Family Medicine*, Vol. 7, No. 2, p. 85, 2010.

6. Schoenberg, P. L. and A. S. David, "Biofeedback for Psychiatric Disorders: A Systematic Review", *Applied Psychophysiology and Biofeedback*, Vol. 39, No. 2, pp. 109–135, 2014.

7. Dorland, W. A. N., *Dorland's Illustrated Medical Dictionary*, WB Saunders, Philadelphia, 1925.

8. Lilly, L. S., *Pathophysiology of Heart Disease: A Collaborative Project of Medical Students and Faculty*, Lippincott Williams & Wilkins, Boston, 2012.

9. Kumar, P. and M. L. Clark, *Kumar and Clark's Clinical Medicine E-Book*, Elsevier Health Sciences, London, 2012.

10. Malik, M., "Heart Rate Variability: Standards of Measurement, Physiological In-

terpretation, and Clinical Use: Task Force of the European Society of Cardiology and the North American Society for Pacing and Electrophysiology", *Annals of Noninvasive Electrocardiology*, Vol. 1, No. 2, pp. 151–181, 1996.

11. McCraty, R. and F. Shaffer, "Heart Rate Variability: New Perspectives on Physiological Mechanisms, Assessment of Self-Regulatory Capacity, and Health Risk", *Global Advances in Health and Medicine*, Vol. 4, No. 1, pp. 46–61, 2015.

12. Shaffer, F. and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms", *Frontiers in Public Health*, Vol. 5, p. 258, 2017.

13. Gevirtz, R., "The Promise of Heart Rate Variability Biofeedback: Evidence-Based Applications", *Biofeedback*, Vol. 41, No. 3, 2013.

14. Jönsson, P., "Respiratory Sinus Arrhythmia as a Function of State Anxiety in Healthy Individuals", *International Journal of Psychophysiology*, Vol. 63, No. 1, pp. 48–54, 2007.

15. Nickel, P. and F. Nachreiner, "Sensitivity and Diagnosticity of the 0.1-Hz Component of Heart Rate Variability as an Indicator of Mental Workload", *Human Factors*, Vol. 45, No. 4, pp. 575–590, 2003.

16. Shelley, K., S. Shelley and C. Lake, "Pulse Oximeter Waveform: Photoelectric Plethysmography", *Clinical Monitoring*, Vol. 2, pp. 420–428, 2001.

17. Al-Jebrni, A. H., B. Chwyl, X. Y. Wang, A. Wong and B. J. Saab, "AI-Enabled Remote and Objective Quantification of Stress at Scale", *Biomedical Signal Processing and Control*, Vol. 59, p. 101929, 2020.

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

19. Umair, M., N. Chalabianloo, C. Sas and C. Ersoy, "HRV and Stress: A Mixed-Methods Approach for Comparison of Wearable Heart Rate Sensors for Biofeedback", *IEEE Access*, Vol. 9, pp. 14005–14024, 2021.

20. Kilpatrick, D. G., "Differential Responsiveness of Two Electrodermal Indices to Psychological Stress and Performance of a Complex Cognitive Task", *Psychophysiology*, Vol. 9, No. 2, pp. 218–226, 1972.

21. Prokasy, W., *Electrodermal Activity in Psychological Research*, Elsevier, New York, 2012.

22. Sanches, P., K. Höök, C. Sas and A. Ståhl, "Ambiguity as a Resource to Inform Proto-Practices: The Case of Skin Conductance", *ACM Transactions on Computer-Human Interaction (TOCHI)*, Vol. 26, No. 4, pp. 1–32, 2019.

23. Lang, P. J., M. M. Bradley and B. N. Cuthbert, "Emotion, Motivation, and Anxiety: Brain Mechanisms and Psychophysiology", *Biological Psychiatry*, Vol. 44, No. 12, pp. 1248–1263, 1998.

24. Martini, F. H. and E. F. Bartholomew, *Essentials of Anatomy & Physiology: Pearson New International Edition*, Pearson Higher Ed, London, 2013.

25. Carlson, N. R., *Physiology of Behavior*, Pearson, Boston, 11 edn., 2012.

26. Boucsein, W., *Electrodermal Activity*, Springer Science & Business Media, Wuppertal, 2012.

27. 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.

28. Greene, S., H. Thapliyal and A. Caban-Holt, "A Survey of Affective Computing for Stress Detection: Evaluating Technologies in Stress Detection for Better

Health", *IEEE Consumer Electronics Magazine*, Vol. 5, No. 4, pp. 44–56, 2016.

29. Alfaras, M., W. Primett, M. Umair, C. Windlin, P. Karpashevich, N. Chalabianloo, D. Bowie, C. Sas, P. Sanches, K. Höök *et al.*, "Biosensing and Actuation, Platforms Coupling Body Input-Output Modalities for Affective Technologies", *Sensors*, Vol. 20, No. 21, p. 5968, 2020.

30. Cacioppo, J. T., L. G. Tassinary and G. Berntson, *Handbook of Psychophysiology*, Cambridge University Press, Cambridge, 2007.

31. Li, C., Q. Chang, J. Zhang and W. Chai, "Effects of Slow Breathing Rate on Heart Rate Variability and Arterial Baroreflex Sensitivity in Essential Hypertension", *Medicine*, Vol. 97, No. 18, 2018.

32. Massaroni, C., A. Nicolò, D. Lo Presti, M. Sacchetti, S. Silvestri and E. Schena, "Contact-Based Methods for Measuring Respiratory Rate", *Sensors*, Vol. 19, No. 4, p. 908, 2019.

33. Kumar, J. S. and P. Bhuvaneswari, "Analysis of Electroencephalography (Eeg) Signals and Its Categorization", *Procedia Engineering*, Vol. 38, pp. 2525–2536, 2012.

34. Sioni, R. and L. Chittaro, "Stress Detection Using Physiological Sensors", *Computer*, Vol. 48, No. 10, pp. 26–33, 2015.

35. Castellano, G., S. D. Villalba and A. Camurri, "Recognising Human Emotions From Body Movement and Gesture Dynamics", *International Conference on Affective Computing and Intelligent Interaction*, pp. 71–82, Berlin, Germany, 2007.

36. Melzer, A., T. Shafir and R. P. Tsachor, "How Do We Recognize Emotion from Movement? Specific Motor Components Contribute to the Recognition of Each Emotion", *Frontiers in Psychology*, Vol. 10, p. 1389, 2019.

37. Hao, Y., J. Budd, M. M. Jackson, M. Sati and S. Soni, "A Visual Feedback Design Based on a Brain-Computer Interface to Assist Users Regulate Their Emotional State", *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, p. 2491–2496, Toronto, Ontario, Canada, 2014.

38. Wilson, G., D. Dobrev and S. A. Brewster, "Hot under the Collar: Mapping Thermal Feedback to Dimensional Models of Emotion", *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4838–4849, San Jose, USA, 2016.

39. Umair, M., C. Sas, N. Chalabianloo and C. Ersoy, "Exploring Personalized Vibrotactile and Thermal Patterns for Affect Regulation", *Designing Interactive Systems Conference*, p. 891–906, New York, USA, 2021.

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

41. Khut, G., "Designing Biofeedback Artworks for Relaxation", *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 3859–3862, San Jose, California, USA, 2016.

42. Jung, H. and E. Stolterman, "Digital Form and Materiality: Propositions for a New Approach to Interaction Design Research", *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, pp. 645–654, Copenhagen, Denmark, 2012.

43. Yu, B., J. Hu, M. Funk and L. M. G. Feijs, "DeLight: Biofeedback through Ambient Light for Stress Intervention and Relaxation Assistance", *Personal and Ubiquitous Computing*, Vol. 22, pp. 787–805, 2018.

44. Wessely, M., T. Tsandilas and W. E. Mackay, "Stretchis: Fabricating Highly Stretchable User Interfaces", *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, p. 697–704, Tokyo, Japan, 2016.

45. Ståhl, A., M. Jonsson, J. Mercurio, A. Karlsson, K. Höök and E.-C. Banka Johnson, "The Soma Mat and Breathing Light", *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, p. 305–308, San Jose, USA, 2016.

46. Keltner, D., K. Oatley and J. M. Jenkins, *Understanding Emotions*, Wiley Hoboken, NJ, Trenton, 2014.

47. Gross, J. J., "The Emerging Field of Emotion Regulation: An Integrative Review", *Review of General Psychology*, Vol. 2, No. 3, pp. 271–299, 1998.

48. Tamir, M., Y. E. Bigman, E. Rhodes, J. Salerno and J. Schreier, "An Expectancy-Value Model of Emotion Regulation: Implications for Motivation, Emotional Experience, and Decision Making", *Emotion*, Vol. 15, No. 1, p. 90, 2015.

49. Gross, J. J., *Emotion Regulation: Conceptual and Empirical Foundations*, The Guilford Press, Washington DC, 2014.

50. Chong, C. S., M. Tsunaka and E. P. Chan, "Effects of Yoga on Stress Management in Healthy Adults: A Systematic Review", *Alternative Therapies in Health and Medicine*, Vol. 17, No. 1, p. 32, 2011.

51. Song, Y. and R. Lindquist, "Effects of Mindfulness-Based Stress Reduction on Depression, Anxiety, Stress and Mindfulness in Korean Nursing Students", *Nurse Education Today*, Vol. 35, No. 1, pp. 86–90, 2015.

52. Arch, J. J., C. R. Ayers, A. Baker, E. Almklov, D. J. Dean and M. G. Craske, "Randomized Clinical Trial of Adapted Mindfulness-Based Stress Reduction versus Group Cognitive Behavioral Therapy for Heterogeneous Anxiety Disorders",

Behaviour Research and Therapy, Vol. 51, No. 4-5, pp. 185–196, 2013.

53. Svetlov, A. S., M. M. Nelson, P. D. Antonenko, J. P. McNamara and R. Bussing, "Commercial Mindfulness Aid Does Not Aid Short-Term Stress Reduction Compared to Unassisted Relaxation", *Heliyon*, Vol. 5, No. 3, p. e01351, 2019.

54. Kanthi, M., A. Puranik and A. Nayak, "Wearable Device for Yogic Breathing with Real-Time Heart Rate and Posture Monitoring", *Journal of Medical Signals and Sensors*, Vol. 11, No. 4, p. 253, 2021.

55. Cheng, P., A. Lucero and J. Buur, "PAUSE: Exploring Mindful Touch Interaction on Smartphones", *Proceedings of the 20th International Academic Mindtrek Conference*, p. 184–191, Tampere, Finland, 2016.

56. Harkness, K. L. and E. P. Hayden, *The Oxford Handbook of Stress and Mental Health*, Oxford University Press, USA, New York, 2020.

57. McEwen, B. S., "Stressed or Stressed Out: What Is the Difference?", *Journal of Psychiatry and Neuroscience*, Vol. 30, No. 5, pp. 315–318, 2005.

58. Wang, M. and K. J. Saudino, "Emotion Regulation and Stress", *Journal of Adult Development*, Vol. 18, No. 2, pp. 95–103, 2011.

59. Griffin, S. M. and S. Howard, "Individual Differences in Emotion Regulation and Cardiovascular Responding to Stress", *Emotion*, Vol. 22, No. 2, p. 331, 2022.

60. Shcherbatykh, Y. V., "Self-Regulation of Autonomic Homeostasis in Emotional Stress", *Human Physiology*, Vol. 26, No. 5, pp. 641–642, 2000.

61. Can, Y. S., H. Iles-Smith, N. Chalabianloo, D. Ekiz, J. Fernández-Álvarez, C. Repetto, G. Riva and C. Ersoy, "How to Relax in Stressful Situations: A Smart Stress Reduction System", *Healthcare*, Vol. 8, p. 100, MDPI, 2020.

62. "Depression: The Treatment and Management of Depression in Adults (Updated Edition)", National Collaborating Centre for Mental Health (UK), British Psychological Society, 2010.

63. Thomée, S., "Mobile Phone Use and Mental Health. A Review of the Research That Takes a Psychological Perspective on Exposure", *International Journal of Environmental Research and Public Health*, Vol. 15, No. 12, p. 2692, 2018.

64. Costa, J., A. T. Adams, M. F. Jung, F. Guimbretière and T. Choudhury, "EmotionCheck: Leveraging Bodily Signals and False Feedback to Regulate Our Emotions", *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, p. 758–769, Heidelberg, Germany, 2016.

65. T Azevedo, R., N. Bennett, A. Bilicki, J. Hooper, F. Markopoulou and M. Tsakiris, "The Calming Effect of a New Wearable Device during the Anticipation of Public Speech", *Scientific Reports*, Vol. 7, No. 1, pp. 1–7, 2017.

66. Costa, J., F. Guimbretière, M. F. Jung and T. Choudhury, "Boostmeup: Improving Cognitive Performance in the Moment by Unobtrusively Regulating Emotions with a Smartwatch", *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 3, No. 2, pp. 1–23, 2019.

67. Paredes, P. E., Y. Zhou, N. A.-H. Hamdan, S. Balters, E. Murnane, W. Ju and J. A. Landay, "Just Breathe: In-Car Interventions for Guided Slow Breathing", *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 2, No. 1, pp. 1–23, 2018.

68. Steffen, P. R., T. Austin, A. DeBarros and T. Brown, "The Impact of Resonance Frequency Breathing on Measures of Heart Rate Variability, Blood Pressure, and Mood", *Frontiers in Public Health*, Vol. 5, p. 222, 2017.

69. Miri, P., E. Jusuf, A. Uusberg, H. Margarit, R. Flory, K. Isbister, K. Marzullo

and J. J. Gross, "Evaluating a Personalizable, Inconspicuous Vibrotactile (PIV) Breathing Pacer for In-The-Moment Affect Regulation", *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, Honolulu, HI, USA, 2020.

70. Jonsson, M., A. Ståhl, J. Mercurio, A. Karlsson, N. Ramani and K. Höök, "The Aesthetics of Heat: Guiding Awareness with Thermal Stimuli", *Proceedings of the TEI'16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction*, pp. 109–117, Eindhoven, Netherlands, 2016.

71. Sanches, P., A. Janson, P. Karpashevich, C. Nadal, C. Qu, C. Daudén Roquet, M. Umair, C. Windlin, G. Doherty, K. Höök and C. Sas, "HCI and Affective Health: Taking Stock of a Decade of Studies and Charting Future Research Directions", *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, p. 1–17, Glasgow, Scotland UK, 2019.

72. Laborde, S., E. Mosley and J. F. Thayer, "Heart Rate Variability and Cardiac Vagal Tone in Psychophysiological Research - Recommendations for Experiment Planning, Data Analysis, and Data Reporting", *Frontiers in Psychology*, Vol. 8, p. 213, 2017.

73. Lane, R. D., K. McRae, E. M. Reiman, K. Chen, G. L. Ahern and J. F. Thayer, "Neural Correlates of Heart Rate Variability during Emotion", *Neuroimage*, Vol. 44, No. 1, pp. 213–222, 2009.

74. Akselrod, S., D. Gordon, F. A. Ubel, D. C. Shannon, A. Berger and R. J. Cohen, "Power Spectrum Analysis of Heart Rate Fluctuation: A Quantitative Probe of Beat-To-Beat Cardiovascular Control", *Science*, Vol. 213, No. 4504, pp. 220–222, 1981.

75. Boonnithi, S. and S. Phongsuphap, "Comparison of Heart Rate Variability Measures for Mental Stress Detection", *Computing in Cardiology*, pp. 85–88,

Hangzhou, China, 2011.

76. Frank, D. L., L. Khorshid, J. F. Kiffer, C. S. Moravec and M. G. McKee, "Biofeedback in Medicine: Who, When, Why and How?", *Mental Health in Family Medicine*, Vol. 7, No. 2, pp. 85–91, 2010.

77. Umair, M., C. Sas and M. H. Latif, "Towards Affective Chronometry: Exploring Smart Materials and Actuators for Real-Time Representations of Changes in Arousal", *Proceedings of the 2019 on Designing Interactive Systems Conference*, DIS '19, p. 1479–1494, San Diego, USA, 2019.

78. McDuff, D., A. Karlson, A. Kapoor, A. Roseway and M. Czerwinski, "AffectAura: An Intelligent System for Emotional Memory", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, p. 849–858, Austin, Texas, USA, 2012.

79. Umair, M., C. Sas and M. Alfaras, "ThermoPixels: Toolkit for Personalizing Arousal-based Interfaces through Hybrid Crafting", *Proceedings of the 2020 on Designing Interactive Systems Conference*, p. 1017–1032, Eindhoven, Netherlands, 2020.

80. Thieme, A., J. Wallace, P. Johnson, J. McCarthy, S. Lindley, P. Wright, P. Olivier and T. D. Meyer, "Design to Promote Mindfulness Practice and Sense of Self for Vulnerable Women in Secure Hospital Services", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, p. 2647–2656, Paris, France, 2013.

81. Vidyarthi, J., B. E. Riecke and D. Gromala, "Sonic Cradle: Designing for an Immersive Experience of Meditation by Connecting Respiration to Music", *Proceedings of the Designing Interactive Systems Conference*, p. 408–417, Newcastle Upon Tyne, United Kingdom, 2012.

82. Lehrer, P. M. and R. Gevirtz, "Heart Rate Variability Biofeedback: How and Why Does It Work?", *Frontiers in Psychology*, Vol. 5, p. 756, 2014.

83. Ghandeharioun, A. and R. Picard, "BrightBeat: Effortlessly Influencing Breathing for Cultivating Calmness and Focus", *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, p. 1624–1631, Denver, Colorado, USA, 2017.

84. Brown, B. B., *Stress and the Art of Biofeedback*, Harper & Row, Oxford, 1977.

85. Schoenberg, P. L. A. and A. S. David, "Biofeedback for Psychiatric Disorders: A Systematic Review", *Applied Psychophysiology and Biofeedback*, Vol. 39, No. 2, pp. 109–135, 2014.

86. Ahani, A., H. Wahbeh, M. Miller, H. Nezamfar, D. Erdogmus and B. Oken, "Change in Physiological Signals during Mindfulness Meditation", *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 1378–1381, San Diego, USA, 2013.

87. Karydis, T., S. Langer, S. L. Foster and A. Mershin, "Identification of Post-Meditation Perceptual States Using Wearable EEG and Self-Calibrating Protocols", *Proceedings of the 11th Pervasive Technologies Related to Assistive Environments Conference*, p. 566–569, Corfu, Greece, 2018.

88. Mason, H., M. Vandoni, G. Debarbieri, E. Codrons, V. Ugargol and L. Bernardi, "Cardiovascular and Respiratory Effect of Yogic Slow Breathing in the Yoga Beginner: What Is the Best Approach?", *Evidence-Based Complementary and Alternative Medicine*, Vol. 2013, 2013.

89. Ingle, R. and R. Awale, "Impact Analysis of Meditation on Physiological Signals", *International Journal on Informatics Visualization*, Vol. 2, pp. 31–36, 2018.

90. Schäfer, A. and J. Vagedes, "How Accurate Is Pulse Rate Variability as an Esti-

mate of Heart Rate Variability?: A Review on Studies Comparing Photoplethysmographic Technology with an Electrocardiogram", *International Journal of Cardiology*, Vol. 166, No. 1, pp. 15–29, 2013.

91. Gil, E., M. Orini, R. Bailon, J. M. Vergara, L. Mainardi and P. Laguna, "Photoplethysmography Pulse Rate Variability as a Surrogate Measurement of Heart Rate Variability during Non-stationary Conditions", *Physiological Measurement*, Vol. 31, No. 9, p. 1271, 2010.

92. Yu, C., Z. Liu, T. McKenna, A. T. Reisner and J. Reifman, "A Method for Automatic Identification of Reliable Heart Rates Calculated from ECG and PPG Waveforms", *Journal of the American Medical Informatics Association*, Vol. 13, No. 3, pp. 309–320, 2006.

93. Mejía-Mejía, E., K. Budidha, T. Y. Abay, J. M. May and P. A. Kyriacou, "Heart Rate Variability (HRV) and Pulse Rate Variability (PRV) for the Assessment of Autonomic Responses", *Frontiers in Physiology*, Vol. 11, p. 779, 2020.

94. Renevey, P., J. Solà, P. Theurillat, M. Bertschi, J. Krauss, D. Andries and C. Sartori, "Validation of a Wrist Monitor for Accurate Estimation of RR Intervals During Sleep", *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 5493–5496, Osaka, Japan, 2013.

95. Lin, W.-H., D. Wu, C. Li, H. Zhang and Y.-T. Zhang, "Comparison of Heart Rate Variability from PPG with That from ECG", *The International Conference on Health Informatics*, pp. 213–215, Cham, Switzerland, 2014.

96. Binsch, O., T. Wabeke and P. Valk, "Comparison of Three Different Physiological Wristband Sensor Systems and Their Applicability for Resilience-and Work Load Monitoring", *BSN 2016 - 13th Annual Body Sensor Networks Conference*, pp. 272–276, San Francisco, CA, USA, 2016.

97. Ge, Z., P. Prasad, N. Costadopoulos, A. Alsadoon, A. Singh and A. Elchouemi, "Evaluating the Accuracy of Wearable Heart Rate Monitors", *IEEE International Conference on Advances in Computing, Communication, and Automation (ICACCA)*, pp. 1–6, Bareilly, India, 2016.

98. Ollander, S., C. Godin, A. Campagne and S. Charbonnier, "A Comparison of Wearable and Stationary Sensors for Stress Detection", *IEEE International Conference on Systems, Man, and Cybernetics, (SMC) - Conference Proceedings*, pp. 4362–4366, Budapest, Hungary, 2016.

99. Bazeley, P., *Qualitative Data Analysis: Practical Strategies*, Sage, London, 2013.

100. Braun, V. and V. Clarke, "Thematic Analysis", *APA Handbook of Research Methods in Psychology, Vol 2: Research Designs: Quantitative, Qualitative, Neuropsychological, And Biological*, pp. 57–71, American Psychological Association, 2012.

101. Fereday, J. and E. Muir-Cochrane, "Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development", *International Journal of Qualitative Methods*, Vol. 5, No. 1, pp. 80–92, 2006.

102. Friese, S., *Qualitative Data Analysis with ATLAS. ti*, SAGE Publications Limited, London, 2019.

103. Bazeley, P. and K. Jackson, *Qualitative Data Analysis With NVivo*, SAGE Publications Limited, London, 2013.

104. Ståhl, A., K. Höök, M. Svensson, A. S. Taylor and M. Combetto, "Experiencing the Affective Diary", *Personal and Ubiquitous Computing*, Vol. 13, No. 5, pp. 365–378, 2008.

105. Sanches, P., K. Höök, E. Vaara, C. Weymann, M. Bylund, P. Ferreira, N. Peira and M. Sjölinder, "Mind the Body! Designing a Mobile Stress Management Appli-

cation Encouraging Personal Reflection", *Proceedings of the 8th ACM Conference on Designing Interactive Systems*, DIS '10, p. 47–56, Aarhus, Denmark, 2010.

106. Ferreira, P., P. Sanches, K. Höök and T. Jaensson, "License to Chill! How to Empower Users to Cope with Stress", *Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges*, p. 123–132, Lund, Sweden, 2008.

107. Bryman, A., "Integrating Quantitative and Qualitative Research: How Is It Done?", *Qualitative Research*, Vol. 6, No. 1, pp. 97–113, 2006.

108. Almalki, S., "Integrating Quantitative and Qualitative Data in Mixed Methods Research–Challenges and Benefits", *Journal of Education and Learning*, Vol. 5, No. 3, pp. 288–296, 2016.

109. Gelo, O., D. Braakmann and G. Benetka, "Quantitative and Qualitative Research: Beyond the Debate", *Integrative Psychological and Behavioral Science*, Vol. 42, No. 3, pp. 266–290, 2008.

110. Adams, A., P. Lunt and P. Cairns, *A Qualititative Approach to HCI Research*, pp. 138–157, Cambridge University Press, Cambridge, 2008.

111. Treiman, D. J., *Quantitative Data Analysis: Doing Social Research to Test Ideas*, John Wiley & Sons, New Jersey, 2014.

112. van Elst, H., "Foundations of Descriptive and Inferential Statistics", *Arxiv Preprint Arxiv:1302.2525*, 2013.

113. Tilton, S., "Review of the State-Trait Anxiety Inventory (STAI)", *News Notes*, Vol. 48, No. 2, pp. 1–3, 2008.

114. Nolen-Hoeksema, S., *Abnormal Psychology*, McGraw Hill Education, New York, 4 edn., 2007.

115. Spielberger, C., S. Sydeman and M. Maruish, "State-Trait Anxiety Inventory and State-Trait Anger Expression Inventory. The Use of Psychological Testing for Treatment Planning and Outcome Assessment", *Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc*, pp. 292–321, 1994.

116. Heeren, A., E. E. Bernstein and R. J. McNally, "Deconstructing Trait Anxiety: A Network Perspective", *Anxiety, Stress, & Coping*, Vol. 31, No. 3, pp. 262–276, 2018.

117. Cohen, S., T. Kamarck and R. Mermelstein, "A Global Measure of Perceived Stress", *Journal of Health and Social Behavior*, Vol. 24, No. 4, pp. 385–396, 1983.

118. Marcus, M. T., P. M. Fine, F. G. Moeller, M. M. Khan, K. Pitts, P. R. Swank and P. Liehr, "Change in Stress Levels Following Mindfulness-Based Stress Reduction in a Therapeutic Community", *Addictive Disorders & Their Treatment*, Vol. 2, No. 3, pp. 63–68, 2003.

119. Carpenter, L. L., A. R. Tyrka, C. J. McDougle, R. T. Malison, M. J. Owens, C. B. Nemeroff and L. H. Price, "Cerebrospinal Fluid Corticotropin-Releasing Factor and Perceived Early-Life Stress in Depressed Patients and Healthy Control Subjects", *Neuropsychopharmacology*, Vol. 29, No. 4, pp. 777–784, 2004.

120. Stein, P. K. and Y. Pu, "Heart Rate Variability, Sleep and Sleep Disorders", *Sleep Medicine Reviews*, Vol. 16, No. 1, pp. 47–66, 2012.

121. Lu, C.-L., X. Zou, W. C. Orr and J. Chen, "Postprandial Changes of Sympathovagal Balance Measured by Heart Rate Variability", *Digestive Diseases and Sciences*, Vol. 44, No. 4, pp. 857–861, 1999.

122. Zimmermann-Viehoff, F., J. Thayer, J. Koenig, C. Herrmann, C. S. Weber and H.-C. Deter, "Short-Term Effects of Espresso Coffee on Heart Rate Variability and Blood Pressure in Habitual and Non-habitual Coffee Consumers - a Randomized

Crossover Study", *Nutritional Neuroscience*, Vol. 19, No. 4, pp. 169–175, 2016.

123. World Medical Association, "World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects", *The Journal of the American Medical Association*, Vol. 310, No. 20, pp. 2191–2194, 2013.

124. Roh, Y., G. Heo and S. E. Whang, "A Survey on Data Collection for Machine Learning: A Big Data-Ai Integration Perspective", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 33, No. 4, pp. 1328–1347, 2019.

125. Willemink, M. J., W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin and M. P. Lungren, "Preparing Medical Imaging Data for Machine Learning", *Radiology*, Vol. 295, No. 1, pp. 4–15, 2020.

126. Kirschbaum, C., K.-M. Pirke and D. H. Hellhammer, "The 'Trier Social Stress Test' a Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting", *Neuropsychobiology*, Vol. 28, No. 1-2, pp. 76–81, 1993.

127. Scarpina, F. and S. Tagini, "The Stroop Color and Word Test", *Frontiers in Psychology*, Vol. 8, p. 557, 2017.

128. Stroop, J. R., "Studies of Interference in Serial Verbal Reactions", *Journal of Experimental Psychology*, Vol. 18, No. 6, p. 643, 1935.

129. Brodal, P., *The Central Nervous System: Structure and Function*, Oxford University Press, Oxford, 2004.

130. Chen, S.-W., J.-W. Liaw, Y.-J. Chang, L.-L. Chuang and C.-T. Chien, "Combined Heart Rate Variability and Dynamic Measures for Quantitatively Characterizing the Cardiac Stress Status during Cycling Exercise", *Computers in Biology and Medicine*, Vol. 63, pp. 133–142, 2015.

131. Hernando, D., A. Hernando, J. A. Casajus, P. Laguna, N. Garatachea and

R. Bailon, "Methodological Framework for Heart Rate Variability Analysis during Exercise: Application to Running and Cycling Stress Testing", *Medical & Biological Engineering & Computing*, Vol. 56, No. 5, pp. 781–794, 2018.

132. Fitch, D. T., J. Sharpnack and S. L. Handy, "Psychological Stress of Bicycling with Traffic: Examining Heart Rate Variability of Bicyclists in Natural Urban Environments", *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 70, pp. 81–97, 2020.

133. Hong, J.-H., J. Ramos and A. K. Dey, "Understanding Physiological Responses to Stressors during Physical Activity", *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, p. 270–279, Pittsburgh, Pennsylvania, USA, 2012.

134. Tarvainen, M. P., J.-P. Niskanen, J. A. Lipponen, P. O. Ranta-Aho and P. A. Karjalainen, "Kubios HRV, Heart Rate Variability Analysis Software", *Computer Methods and Programs in Biomedicine*, Vol. 113, No. 1, pp. 210–220, 2014.

135. Pan, J. and W. J. Tompkins, "A Real-Time QRS Detection Algorithm", *IEEE Transactions on Biomedical Engineering*, Vol. 32, No. 3, pp. 230–236, 1985.

136. Lipponen, J. A. and M. P. Tarvainen, "A Robust Algorithm for Heart Rate Variability Time Series Artefact Correction Using Novel Beat Classification", *Journal of Medical Engineering & Technology*, Vol. 43, No. 3, pp. 173–181, 2019.

137. Can, Y. S., B. Arnrich and C. Ersoy, "Stress Detection in Daily Life Scenarios Using Smart Phones and Wearable Sensors: A Survey", *Journal of Biomedical Informatics*, Vol. 92, p. 103139, 2019.

138. Taylor, S., N. Jaques, W. Chen, S. Fedor, A. Sano and R. Picard, "Automatic Identification of Artifacts in Electrodermal Activity Data", *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1934–1937, Milan, Italy, 2015.

139. Makowski, D., T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel and S. H. A. Chen, "NeuroKit2: A Python Toolbox for Neurophysiological Signal Processing", *Behavior Research Methods*, Vol. 53, No. 4, pp. 1689–1696, 2021.

140. 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, 2015.

141. Castaldo, R., L. Montesinos, P. Melillo, C. James and L. Pecchia, "Ultra-Short Term Hrv Features as Surrogates of Short Term HRV: A Case Study on Mental Stress Detection in Real Life", *BMC Medical Informatics and Decision Making*, Vol. 19, No. 1, pp. 1–13, 2019.

142. Baek, H. J., C.-H. Cho, J. Cho and J.-M. Woo, "Reliability of Ultra-Short-Term Analysis as a Surrogate of Standard 5-min Analysis of Heart Rate Variability", *Telemedicine and e-Health*, Vol. 21, No. 5, pp. 404–414, 2015.

143. Pernice, R., M. Javorka, J. Krohova, B. Czippelova, Z. Turianikova, A. Busacca, L. Faes *et al.*, "Comparison of Short-Term Heart Rate Variability Indexes Evaluated through Electrocardiographic and Continuous Blood Pressure Monitoring", *Medical & Biological Engineering & Computing*, Vol. 57, No. 6, pp. 1247–1263, 2019.

144. Hernando, D., S. Roca, J. Sancho, Á. Alesanco and R. Bailón, "Validation of the Apple Watch for Heart Rate Variability Measurements during Relax and Mental Stress in Healthy Subjects", *Sensors*, Vol. 18, No. 8, p. 2619, 2018.

145. Vescio, B., M. Salsone, A. Gambardella and A. Quattrone, "Comparison between Electrocardiographic and Earlobe Pulse Photoplethysmographic Detection for Evaluating Heart Rate Variability in Healthy Subjects in Short-and Long-Term Recordings", *Sensors*, Vol. 18, No. 3, p. 844, 2018.

146. Barrios, L., P. Oldrati, S. Santini and A. Lutterotti, "Evaluating the Accuracy of Heart Rate Sensors Based on Photoplethysmography for In-the-Wild Analysis", *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, p. 251–261, Trento, Italy, 2019.

147. Gilgen-Ammann, R., T. Schweizer and T. Wyss, "RR Interval Signal Quality of a Heart Rate Monitor and an ECG Holter at Rest and during Exercise", *European Journal of Applied Physiology*, Vol. 119, No. 7, pp. 1525–1532, 2019.

148. Plews, D. J., B. Scott, M. Altini, M. Wood, A. E. Kilding and P. B. Laursen, "Comparison of Heart-Rate-Variability Recording with Smartphone Photoplethysmography, Polar H7 Chest Strap, and Electrocardiography", *International Journal of Sports Physiology and Performance*, Vol. 12, No. 10, pp. 1324–1328, 2017.

149. Menghini, L., E. Gianfranchi, N. Cellini, E. Patron, M. Tagliabue and M. Sarlo, "Stressing the Accuracy: Wrist-Worn Wearable Sensor Validation over Different Conditions", *Psychophysiology*, Vol. 56, No. 11, p. e13441, 2019.

150. Regalia, G., F. Onorati, M. Lai, C. Caborni and R. W. Picard, "Multimodal Wrist-Worn Devices for Seizure Detection and Advancing Research: Focus on the Empatica Wristbands", *Epilepsy Research*, Vol. 153, pp. 79 – 82, 2019.

151. 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.

152. Montesinos, V., F. Dell'Agnola, A. Arza, A. Aminifar and D. Atienza, "Multi-Modal Acute Stress Recognition Using Off-The-Shelf Wearable Devices", *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2196–2201, Berlin, Germany, 2019.

153. Matsumoto, Y., T. Mizuno, K. Mito and N. Itakura, "Mental Stress Evaluation Method Using Photoplethysmographic Amplitudes Obtained from a Smartwatch", *International Conference on Human-Computer Interaction*, pp. 357–362, Springer, Washington DC, USA, 2021.

154. Jaiswal, D., A. Chowdhury, D. Chatterjee and R. Gavas, "Unobtrusive Smart-Watch Based Approach for Assessing Mental Workload", *2019 IEEE Region 10 Symposium (TENSYMP)*, pp. 304–309, Kolkata, India, 2019.

155. Shcherbina, A., C. M. Mattsson, D. Waggott, H. Salisbury, J. W. Christle, T. Hastie, M. T. Wheeler and E. A. Ashley, "Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort", *Journal of Personalized Medicine*, Vol. 7, No. 2, p. 3, 2017.

156. Weston, E., P. Le and W. S. Marras, "A Biomechanical and Physiological Study of Office Seat and Tablet Device Interaction", *Applied Ergonomics*, Vol. 62, pp. 83 – 93, 2017.

157. Ottaviani, C., L. Shahabi, M. Tarvainen, I. Cook, M. Abrams and D. Shapiro, "Cognitive, Behavioral, and Autonomic Correlates of Mind Wandering and Perseverative Cognition in Major Depression", *Frontiers in Neuroscience*, Vol. 8, p. 433, 2015.

158. Crespo-Ruiz, B., S. Rivas-Galan, C. Fernandez-Vega, C. Crespo-Ruiz and L. Maicas-Perez, "Executive Stress Management: Physiological Load of Stress and Recovery in Executives on Workdays", *International Journal of Environmental Research and Public Health*, Vol. 15, No. 12, p. 2847, 2018.

159. Föhr, T., A. Tolvanen, T. Myllymäki, E. Järvelä-Reijonen, K. Peuhkuri, S. Rantala, M. Kolehmainen, R. Korpela, R. Lappalainen, M. Ermes *et al.*, "Physical Activity, Heart Rate Variability–Based Stress and Recovery, and Subjective Stress during a 9-Month Study Period", *Scandinavian Journal of Medicine &*

Science in Sports, Vol. 27, No. 6, pp. 612–621, 2017.

160. Ahmed, N. and Y. Zhu, "Early Detection of Atrial Fibrillation Based on ECG Signals", *Bioengineering*, Vol. 7, No. 1, p. 16, 2020.

161. Batista, D., H. P. da Silva, A. Fred, C. Moreira, M. Reis and H. A. Ferreira, "Benchmarking of the Bitalino Biomedical Toolkit against an Established Gold Standard", *Healthcare Technology Letters*, Vol. 6, No. 2, pp. 32–36, 2019.

162. Hasanbasic, A., M. Spahic, D. Bosnjic, V. Mesic, O. Jahic *et al.*, "Recognition of Stress Levels among Students with Wearable Sensors", *IEEE 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, pp. 1–4, East Sarajevo, Bosnia and Herzegovina, 2019.

163. Günaydin, Ö. and R. B. Arslan, "Stress Level Detection Using Physiological Sensors", *IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 509–512, Cincinnati, OH, USA, 2020.

164. Speer, K. E., S. Semple, N. Naumovski and A. J. McKune, "Measuring Heart Rate Variability Using Commercially Available Devices in Healthy Children: A Validity and Reliability Study", *European Journal of Investigation in Health, Psychology and Education*, Vol. 10, No. 1, pp. 390–404, 2020.

165. Umair, M., N. Chalabianloo, C. Sas and C. Ersoy, "A Comparison of Wearable Heart Rate Sensors for HRV Biofeedback in the Wild: An Ethnographic Study", *25th annual international CyberPsychology, CyberTherapy & Social Networking Conference*, Milan, Italy, 2020.

166. Chen, C., C. Li, C.-W. Tsai and X. Deng, "Evaluation of Mental Stress and Heart Rate Variability Derived from Wrist-Based Photoplethysmography", *IEEE Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*, pp. 65–68, Okinawa, Japan, 2019.

167. Nwaogu, J. M. and A. P. Chan, "Work-Related Stress, Psychophysiological Strain, and Recovery among On-Site Construction Personnel", *Automation in Construction*, Vol. 125, p. 103629, 2021.

168. Mishra, V., S. Sen, G. Chen, T. Hao, J. Rogers, C.-H. Chen and D. Kotz, "Evaluating the Reproducibility of Physiological Stress Detection Models", *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 4, No. 4, pp. 1–29, 2020.

169. Miranda, D., M. Calderón and J. Favela, "Anxiety Detection Using Wearable Monitoring", *Proceedings of the 5th Mexican Conference on Human-Computer Interaction*, p. 34–41, Oaxaca, Mexico, 2014.

170. Bakhchina, A. V., K. R. Arutyunova, A. A. Sozinov, A. V. Demidovsky and Y. I. Alexandrov, "Sample Entropy of the Heart Rate Reflects Properties of the System Organization of Behaviour", *Entropy*, Vol. 20, No. 6, p. 449, 2018.

171. Chernigovskaya, T. V., S. B. Parin, I. S. Parina, A. A. Konina, D. K. Urikh, Y. O. Yachmonina, M. A. Chernova and S. A. Polevaya, "Simultaneous Interpreting and Stress: Pilot Experiment", *International Journal of Psychophysiology*, Vol. 108, p. 165, 2016.

172. Mark, G., S. T. Iqbal, M. Czerwinski, P. Johns, A. Sano and Y. Lutchyn, "Email Duration, Batching and Self-Interruption: Patterns of Email Use on Productivity and Stress", *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 1717–1728, San Jose, California, USA, 2016.

173. Afulani, P. A., L. Ongeri, J. Kinyua, M. Temmerman, W. B. Mendes and S. J. Weiss, "Psychological and Physiological Stress and Burnout among Maternity Providers in a Rural County in Kenya: Individual and Situational Predictors", *BMC Public Health*, Vol. 21, No. 1, pp. 1–16, 2021.

174. Hettiarachchi, I. T., S. Hanoun, D. Nahavandi and S. Nahavandi, "Validation of Polar OH1 Optical Heart Rate Sensor for Moderate and High Intensity Physical Activities", *Public Library of Science One*, Vol. 14, No. 5, 2019.

175. Determan, J., M. A. Akers, T. Albright, B. Browning, C. Martin-Dunlop, P. Archibald and V. Caruolo, "The Impact of Biophilic Learning Spaces on Student Success", `https://cgdarch.com/wp-content/uploads/2019/12/The-Impact-of-Biophilic-Learning-Spaces-on-Student-Success.pdf`, accessed on September 28, 2022.

176. White, H., *The Effects of Mindfulness Exercises Derived from Acceptance and Commitment Therapy during Recovery from Work-Related Stress*, M.S. Thesis, California State University, 2019.

177. "Accuracy of Firstbeat Bodyguard 2 Beat-To-Beat Heart Rate Monitor", `https://www.firstbeat.com/wp-content/uploads/2015/10/white_paper_bodyguard2_final.pdf`, accessed on September 28, 2022.

178. "Empatica, Medical Devices, AI and Algorithms for Remote Patient Monitoring", `https://www.empatica.com/research/e4/`, accessed on September 28, 2022.

179. "Own Your Temperature, Feel Colder or Warmer", `https://embrlabs.com/`, accessed on September 28, 2022.

180. "Feel Calm and Focused, Naturally", `https://feeldoppel.com/`, accessed on September 28, 2022.

181. Smith, M. J., K. Warren, D. Cohen-Tanugi, S. Shames, K. Sprehn, J. L. Schwartz, H. Zhang and E. Arens, "Augmenting Smart Buildings and Autonomous Vehicles with Wearable Thermal Technology", *International Conference on Human-Computer Interaction*, pp. 550–561, Cham, Switzerland, 2017.

182. Matlin, M. W. and H. J. Foley, *Sensation and Perception*, Allyn & Bacon, Boston,

1992.

183. Wilson, G., M. Halvey, S. A. Brewster and S. A. Hughes, "Some like It Hot: Thermal Feedback for Mobile Devices", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, p. 2555–2564, Vancouver, Canada, 2011.

184. Jones, L., "Thermal Touch", *Scholarpedia*, Vol. 4, No. 5, p. 7955, 2009.

185. Greenwald, A. G., "Within-Subjects Designs: To Use or Not to Use?", *Psychological Bulletin*, Vol. 83, No. 2, p. 314, 1976.

186. Hollander, M., D. A. Wolfe and E. Chicken, *Nonparametric Statistical Methods*, Vol. 751, John Wiley & Sons, Hoboken, 2013.

187. Neuhäuser, M. and F. Bretz, "Nonparametric All-Pairs Multiple Comparisons", *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, Vol. 43, No. 5, pp. 571–580, 2001.

188. Hartas, D., *Educational Research and Inquiry: Qualitative and Quantitative Approaches*, Bloomsbury Publishing, London, 2015.

189. Douglas, C. E. and F. A. Michael, "On Distribution-Free Multiple Comparisons in the One-Way Analysis of Variance", *Communications in Statistics-Theory and Methods*, Vol. 20, No. 1, pp. 127–139, 1991.

190. Dolgun, A. and H. Demirhan, "Performance of Nonparametric Multiple Comparison Tests under Heteroscedasticity, Dependency, and Skewed Error Distribution", *Communications in Statistics-Simulation and Computation*, Vol. 46, No. 7, pp. 5166–5183, 2017.

191. Dmitrienko, A., C. Chuang-Stein and R. B. D'Agostino Sr, *Pharmaceutical Statistics Using Sas: A Practical Guide*, SAS Institute, Cary, 2007.

192. Juneau, P., "Simultaneous Nonparametric Inference in a One-Way Layout Using the Sas System", *Proceedings of the PharamSUG 2004 Annual Meeting*, p. 112, San Diego, USA, 2004.

193. Terpilowski, M., "Scikit-Posthocs: Pairwise Multiple Comparison Tests in Python", *The Journal of Open Source Software*, Vol. 4, No. 36, p. 1169, 2019.

194. Rauh, R., R. Limley, R.-D. Bauer, M. Radespiel-Troger and M. Mueck-Weymann, "Comparison of Heart Rate Variability and Pulse Rate Variability Detected with Photoplethysmography", *Saratov Fall Meeting: Optical Technologies in Biophysics and Medicine V*, pp. 115–126, Saratov, Russia, 2004.

195. Jo, E., K. Lewis, D. Directo, M. J. Kim and B. A. Dolezal, "Validation of Biofeedback Wearables for Photoplethysmographic Heart Rate Tracking", *Journal of Sports Science & Medicine*, Vol. 15, No. 3, p. 540, 2016.

196. Ehmen, H., M. Haesner, I. Steinke, M. Dorn, M. Gövercin and E. Steinhagen-Thiessen, "Comparison of Four Different Mobile Devices for Measuring Heart Rate and Ecg With Respect to Aspects of Usability and Acceptance by Older People", *Applied Ergonomics*, Vol. 43, No. 3, pp. 582–587, 2012.

197. McCarthy, C., N. Pradhan, C. Redpath and A. Adler, "Validation of the Empatica E4 Wristband", *2016 IEEE EMBS International Student Conference (ISC)*, pp. 1–4, Ottawa, Canada, 2016.

198. Kutt, K., W. Binek, P. Misiak, G. J. Nalepa and S. Bobek, "Towards the Development of Sensor Platform for Processing Physiological Data from Wearable Sensors", *International Conference on Artificial Intelligence and Soft Computing*, pp. 168–178, Zakopane, Poland, 2018.

199. Bhowmik, T., J. Dey and V. N. Tiwari, "A Novel Method for Accurate Estimation of HRV from Smartwatch PPG Signals", *2017 39th Annual International*

Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 109–112, Jeju, Korea, 2017.

200. Föhr, T., A. Tolvanen, T. Myllymäki, E. Järvelä-Reijonen, S. Rantala, R. Korpela, K. Peuhkuri, M. Kolehmainen, S. Puttonen, R. Lappalainen *et al.*, "Subjective Stress, Objective Heart Rate Variability-Based Stress, and Recovery on Workdays among Overweight and Psychologically Distressed Individuals: A Cross-Sectional Study", *Journal of Occupational Medicine and Toxicology*, Vol. 10, No. 1, p. 39, 2015.

201. Can, Y. S., N. Chalabianloo, D. Ekiz, J. Fernández-Álvarez, C. Repetto, G. Riva, H. Iles-Smith and C. Ersoy, "Real-Life Stress Level Monitoring using Smart Bands in the Light of Contextual Information", *IEEE Sensors Journal*, Vol. 20, pp. 8721–8730, 2020.

202. Simonnet, M. and B. Gourvennec, "Heart Rate Sensors Acceptability: Data Reliability vs. Ease of Use", *IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 94–98, San Francisco, CA, USA, 2016.

203. Ranganathan, S., K. Nakai and C. Schonbach, *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, Elsevier, New York, 2018.

204. Van Lier, H. G., M. E. Pieterse, A. Garde, M. G. Postel, H. A. de Haan, M. M. Vollenbroek-Hutten, J. M. Schraagen and M. L. Noordzij, "A Standardized Validity Assessment Protocol for Physiological Signals from Wearable Technology: Methodological Underpinnings and an Application to the E4 Biosensor", *Behavior Research Methods*, Vol. 52, pp. 607–629, 2019.

205. Bulte, C. S., S. W. Keet, C. Boer and R. A. Bouwman, "Level of Agreement between Heart Rate Variability and Pulse Rate Variability in Healthy Individuals", *European Journal of Anaesthesiology (EJA)*, Vol. 28, No. 1, pp. 34–38, 2011.

206. Schrödl, E., S. Kampusch, B. D. Razlighi, V. H. Le, J. C. Széles and E. Kaniusas, "Feasibility of Pulse Rate Variability as Feedback in Closed-Loop Percutaneous Auricular Vagus Nerve Stimulation", *Vibroengineering PROCEDIA*, Vol. 26, pp. 35–39, 2019.

207. Tonacci, A., L. Billeci, E. Burrai, F. Sansone and R. Conte, "Comparative Evaluation of the Autonomic Response to Cognitive and Sensory Stimulations through Wearable Sensors", *Sensors*, Vol. 19, No. 21, p. 4661, 2019.

208. Giavarina, D., "Understanding Bland Altman Analysis", *Biochemia Medica*, Vol. 25, No. 2, pp. 141–151, 2015.

209. Altman, D. G. and J. M. Bland, "Measurement in Medicine: The Analysis of Method Comparison Studies", *Journal of the Royal Statistical Society: Series D (The Statistician)*, Vol. 32, No. 3, pp. 307–317, 1983.

210. Bland, J. M. and D. G. Altman, "Measuring Agreement in Method Comparison Studies", *Statistical Methods in Medical Research*, Vol. 8, No. 2, pp. 135–160, 1999.

211. Altman, D. G., J Martin, "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement", *International Journal of Nursing Studies*, Vol. 47, No. 8, pp. 931–936, 2010.

212. Bland, J. M., D. G. Altman and D. S. Warner, "Agreed Statistics: Measurement Method Comparison", *The Journal of the American Society of Anesthesiologists*, Vol. 116, No. 1, pp. 182–185, 2012.

213. "Cardiac Monitors, Heart Rate Meters, and Alarms", American National Standard (ANSI/AAMI EC13: 2002), Arlington, VA, 2002.

214. Lykken, D., R. Rose, B. Luther and M. Maley, "Correcting Psychophysiological Measures for Individual Differences in Range", *Psychological Bulletin*, Vol. 66,

No. 6, p. 481, 1966.

215. Kursa, M. B. and W. R. Rudnicki, "Feature Selection With the Boruta Package", *Journal of Statistical Software*, Vol. 36, pp. 1–13, 2010.

216. Cawley, G. C. and N. L. Talbot, "On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation", *The Journal of Machine Learning Research*, Vol. 11, pp. 2079–2107, 2010.

217. Varma, S. and R. Simon, "Bias in Error Estimation When Using Cross-Validation for Model Selection", *BMC Bioinformatics*, Vol. 7, No. 1, pp. 1–8, 2006.

218. Krstajic, D., L. J. Buturovic, D. E. Leahy and S. Thomas, "Cross-Validation Pitfalls When Selecting and Assessing Regression and Classification Models", *Journal of Cheminformatics*, Vol. 6, No. 1, pp. 1–15, 2014.

219. Breiman, L., "Random Forests", *Machine Learning*, Vol. 45, No. 1, pp. 5–32, 2001.

220. Friedman, J., T. Hastie, R. Tibshirani *et al.*, *The Elements of Statistical Learning*, Vol. 1, Springer Series in Statistics, New York, 2001.

221. Geurts, P., D. Ernst and L. Wehenkel, "Extremely Randomized Trees", *Machine Learning*, Vol. 63, No. 1, pp. 3–42, 2006.

222. Natekin, A. and A. Knoll, "Gradient Boosting Machines, a Tutorial", *Frontiers in Neurorobotics*, Vol. 7, p. 21, 2013.

223. Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, "Lightgbm: A Highly Efficient Gradient Boosting Decision Tree", *Advances in Neural Information Processing Systems*, Vol. 30, pp. 3146–3154, 2017.

224. Raschka, S., "Model Evaluation, Model Selection, and Algorithm Selection in

Machine Learning", *arXiv:1811.12808*, 2018.

225. Dinno, A., "Nonparametric Pairwise Multiple Comparisons in Independent Groups Using Dunn's Test", *The Stata Journal*, Vol. 15, No. 1, pp. 292–300, 2015.

226. Lundberg, S. M. and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions", *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–4777, Long Beach, California USA, 2017.

227. Ferreira, L. A., F. G. Guimarães and R. Silva, "Applying Genetic Programming to Improve Interpretability in Machine Learning Models", *2020 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8, Glasgow, United Kingdom, 2020.

228. Deo, R. C., "Machine Learning in Medicine", *Circulation*, Vol. 132, No. 20, pp. 1920–1930, 2015.

229. Baniecki, H., W. Kretowicz, P. Piatyszek, J. Wisniewski and P. Biecek, "Dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python", *Journal of Machine Learning Research*, Vol. 22, No. 214, pp. 1–7, 2021.

230. Ribeiro, M. T., S. Singh and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier", *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 1135–1144, San Francisco, California, USA, 2016.

231. Mazzanti, S., "SHAP Values Explained Exactly How You Wished Someone Explained to You", *Towards Data Science*, Vol. 3, p. 2020, 2020.

232. Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee, "From Local Explanations to Global Understanding with Explainable AI for Trees", *Nature Machine Intelligence*, Vol. 2, No. 1, pp. 56–67, 2020.

APPENDIX A: USE OF COPYRIGHTED MATERIAL

All figures and tables in Chapters 1, 2, and 8 are illustrated and plotted by Niaz Chalebianloo and have not previously been used in any other article or sources. The rest of the figures used in this thesis are also all illustrated by Niaz Chalabianloo and have previously been used in our papers in [19], [39], and [61]. We have published all these papers as open access articles distributed under the Creative Commons Attribution License. The copyright status for these papers is CC BY, meaning that the journal is not the copyright holder of these materials. Under CC BY, reuse for commercial purposes or to create derivative works is permitted, and readers can copy and redistribute the material in any medium or format.