Application Level Performance Evaluation of Wearable Devices for Stress Classification with Explainable AI

Niaz Chalabianloo[a,*], Yekta Said Can[b], Muhammad Umair[c], Corina Sas[d] and Cem Ersoy[a]

[a]Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey
[b]Faculty of Applied Computer Science, Universität Augsburg, Augsburg, Germany
[c]Newcastle University, Newcastle Upon Tyne, United Kingdom
[d]School of Computing and Communications, Lancaster University, Lancaster, United Kingdom

ARTICLE INFO

Keywords:
HRV
EDA
XAI
Stress Detection
Wearable Sensors
Affective Computing

ABSTRACT

Stress has become one of the most prominent problems of modern societies and a key contributor to major health issues. Dealing with stress effectively requires detecting it in real-time, informing the user, and giving instructions on how to manage it. Over the past few years, wearable devices equipped with biosensors that can be utilized for stress detection have become increasingly popular. Since they come with various designs and technologies and acquire biosignals from different body locations, choosing a suitable device for a particular application has become a challenge for researchers and end-users. This study compares seven common wearable biosensors for stress detection applications. This was accomplished by collecting physiological sensor data during Baseline, Stress, Recovery, and Cycling sessions from 32 participants. Machine learning algorithms were used to classify four stress classes, and the results obtained from all wearables were compared. Following this, a state-of-the-art explainable artificial intelligence method was employed to clarify our models' predictions and investigate the influence different features have on the models' outputs. Despite the results showing that ECG wearables perform slightly better than the rest of the devices, adding a second biosignal (EDA) improved the results significantly, tipping the balance toward multisensor wearables. Finally, we concluded that although the output results of each model can be affected by various factors, in most cases, there is no significant difference in the accuracy of stress detection by different wearables. However, the decision to select a particular wearable for stress detection applications must be made carefully considering the trade-off between the users' expectations and preferences and the pros and cons of each device.

1. Introduction

Stress has become a major underlying cause for health-related problems, exposing people to varying degrees of mental and physical ailments [1] with increased societal cost. According to a report by the American Psychological Association, it has been estimated that stress-related health problems and their consequences cost the United States alone $300 billion annually [2].

Nowadays, one of the most effective methods of dealing with stress is developing the ability to self-regulate one's physiological and mental state [3, 4, 5]. Self-regulation of stress can be achieved using biofeedback since it helps users become self-aware of their stress levels and subsequently perform the necessary interventions [4]. Sufficient self-regulation of daily life stresses is achievable using cost-effective interventions with short durations [6]. For instance, relaxation is regarded as the most practical among different stress management interventions since it is the most inexpensive and most straightforward to

*Corresponding author at: Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey
niaz.chalabianloo@boun.edu.tr (N. Chalabianloo)
ORCID(s): 0000-0001-7228-4725 (N. Chalabianloo); 0000-0002-6614-0183 (Y.S. Can); 0000-0002-0017-9827 (M. Umair); 0000-0001-9297-9612 (C. Sas); 0000-0001-7632-7067 (C. Ersoy)

implement [6, 7]. Other methods such as mobile applications for practicing mindfulness also show promising results for effective stress reduction [8, 9]. Accordingly, it can be concluded that in an efficient stress management system, the first and most crucial step is to detect the occurrence and levels of stress for informing the user.

One of the most common, reliable, and easy to access methods for detecting stress is the use of physiological signals, characterizing changes in physiological arousal [10]. In a rudimentary wearable sensor not customized for a specific application, only the physiological signals and their fluctuations are possible to observe. For instance, the HRV signal and its features can be utilized to detect mental stress. However, the end-user may not comprehend what these signals reveal, and the presence of an expert will be required for meaningful interpretation necessary to detect signs of stress. Nowadays, machine learning (ML) methods are used in designing stress detection mechanisms for everyday life. For making these mechanisms capable of interpreting the data and recognizing users' stress levels, ML algorithms are trained using previous data.

In recent decades, collecting physiological signals from the human body has often required cumbersome laboratory equipment. Nevertheless, due to significant advancements in Micro-Electro-Mechanical Systems (MEMS), the sizes of these sensors have considerably shrunk while preserving an adequate level of measurements' quality [11]. Much research aimed to develop and improve the ubiquitous and wearable affective computing technologies in order to help improve people's mental well-being [12]. Wearable sensors are offered in the form of single-purpose devices or as add-ons to smartphones and smartwatches. As a result, people can easily carry a variety of sensors in their pockets, on their wrists, or arms to record and examine their body's physiological signals. For instance, nearly all low-cost smartwatches are equipped with Photoplethysmography (PPG) sensors to measure cardiovascular signals, while more expensive ones include sensors such as Accelerometer (ACC), Electrocardiography (ECG), Electrodermal Activity (EDA), and those for body temperature. While biosignal acquisition is a promising feature of personal wearable devices, there are some disadvantages to consider as well. For instance, they have to fairly divide their battery capacity into multiple purposes, and this energy optimization yields a trade-off between availability and performance. On the contrary, medical-grade devices, or even wearable devices with larger sizes that have no battery restrictions and are solely designed for a single purpose, can record continuously without energy restrictions and with a higher sampling rate, resulting in higher accuracy and superior data quality.

Much previous research on the validity of wearables have concluded that their measurement accuracy is lower than that of medical-grade devices [13, 14, 15]. Despite this, the demand for wearable devices continues to grow due to multiple factors such as their unobtrusiveness and ease of use [16]. Besides, the measurement quality of these devices continues to increase with the introduction of newer generations equipped with enhanced technologies and capabilities [15, 17, 18]. It is expected that by 2023, there will be more than 330 million people who use at least one wearable device, and the market is expanding every year [19]. However, the wide variety of wearable gadgets on the market can make it difficult for end-users and researchers to choose the right ones for stress measurement applications. In order to choose the best available options, it is necessary to make a comprehensive comparison between different wearable devices.

Since the signals collected by biosensors are the building blocks in the biofeedback process, in a previous study, we conducted a comprehensive comparative study by comparing the physiological data quality collected simultaneously from several devices [20]. The HRV features were compared using correlation and agreement analysis in several conditions. However, all comparisons were made only at the feature level, and neither device was investigated under a specific application (e.g., stress measurement). Findings showed that single-purpose ECG chest bands exhibited relatively superior data quality compared to smart and non-smart wristbands fitted with PPG sensors. One of the main reasons for this difference in data quality was the extreme adverse effects of various noises, such as physical movements on devices with PPG sensors. It was also observed that by applying proper noise reduction methods, output data from PPG sensors could become similar to and more correlated with ECG sensor data. Following the quantitative analysis, in order to

investigate the usability and users' preferences, the previous study also involved a qualitative evaluation. By interviewing users of a wide range of smartwatches, wristbands, chest bands, and other specialized wearables, we found that most users favored smaller devices, preferably with screens, and in a format that could be worn on the wrist. Thus, in general, and regardless of a specific type of application, factors such as wearability, ease of use, and comfort can tip the balance in favor of choosing devices with lower data quality.

While this previous paper focused only on data quality [20], the present paper examines the effects of using the data produced by each device in a particular application (i.e., stress level detection). In this paper, we report an experiment where participants were first subjected to mental stress followed by intense physical activity. We wanted to find out whether devices equipped with ECG sensors still function much better and retain their superiority, or are we going to witness a close competition? For this purpose, we studied the difference in stress level detection quality between these devices in terms of accuracy and precision by making higher-level (application level) comparisons using ML algorithms. Features extracted from three types of biosignals (i.e., ECG, PPG, and EDA) of a comprehensive set of wearables (seven devices) were used for training and testing the ML models.

Moreover, a state-of-the-art explainable AI (XAI) method was utilized to investigate various factors, including the number and importance of features, data preprocessing methods, effects of different models, and the impacts of all these factors on the models' outputs. Furthermore, we inspected the importance of preventing data leakage and examined whether different features exhibit similar behaviors under the influence of various selections of classifiers and scaling algorithms. To the best of our knowledge, this study is the first of its kind, which compares and evaluates a comprehensive set of wearable devices for identifying stress and physical activity. The main contributions of this study are (i) to help end-users and researchers make better decisions in choosing wearable devices for stress level detection, and (ii) to introduce the advantages of using XAI in affective computing for better comprehension of the features' importance and their impact on the selection of a particular class by ML models.

2. Background and Related Works

Stress has become an inseparable part of modern life. Many factors, such as highly competitive work environments, heavy workloads, time pressure, and financial conditions, contribute to excessive stress among individuals. The non-formal definition of stress is the human body's response to any challenging or dangerous situation. When people are stimulated with such stressors, they check their inner resources to handle them [11]. No matter how difficult the situation is, if people feel they have enough resources, they will not feel stressed. Otherwise, their mental stress levels will increase. There are different stress types when the duration and evaluation reference are considered [11, 21]. Long-term stress is called chronic, whereas short-term stress is named acute stress. When the reference is taken into consideration, the body's physiological reaction to a stressor can be named physiological stress. The context (i.e., stressed and relaxed) of individuals is used as a reference in this type of stress. On the other hand, mental appraisal and interpretation of stressful situations is called perceived stress. Self-report questionnaires are used for measuring perceived stress.

Traditionally, stress has been measured by analyzing self-report questionnaires or being interviewed by a psychologist. However, with the advancement of wearable technologies, automatic stress monitoring systems have been developed. They collect physiological information related to the intensity of stress and experience of a subject. The related signals could be listed as Heart Rate Variability (HRV), Brain Activity, Electro-Dermal Activity (EDA), also known as Galvanic Skin Response (GSR), Blood Pressure (BP), Skin Temperature (ST), Muscle Activity, Respiration, Blood Volume Pulse (BVP).

One of the most significant signals for determining stress levels is heart activity since the autonomic nervous system influences the heart directly. The most commonly extracted feature of heart activity is Heart Rate Variability (HRV) which can be defined as the change in the time interval between successive heartbeats. Under emotional arousal, sinusoidal heart rhythms are corrupted, and they become more irregular and faster,

and this effect could be observed from HRV very clearly. That is why we selected HRV for physiological biofeedback. Two leading sensor technology can measure heart activity: ECG and PPG namely and we used both technologies in this study.

There are different levels of validity assessment of wearable devices [22]. The first level is signal level, the most straightforward comparison that evaluates the ability of a device to generate the same raw data as the reference device. Correlation is generally computed for this level of validity assessment. The second one is the feature (parameter) level: it determines whether a device delivers physiological features similar to the reference device. The Bland–Altman plot is the most commonly applied technique for this level. Lastly, the application (event) level compares devices with the reference device on the capability to recognize an application (event). Usually, ML algorithms are used with both devices, and metrics such as accuracy, precision, recall, and F_1 score are then compared.

In the literature, wearable devices are compared with medical-grade ones, especially at the feature level. These studies demonstrate that wearable PPG sensors are catching up with medical-grade ECG sensors for extracting HRV features. Bolanos et al. [23] reported that there is an excellent agreement between different measures of HRV signals acquired from both sensors, and the presented results provide potential support for the idea of using PPGs instead of ECGs in HRV signal derivation and analysis in ambulatory cardiac monitoring of healthy individuals. Selvaraj et al. [24] also inferred that HRV features can be reliably estimated from the PPG. Jeyhani et al. [25] compared PPG-based wearable devices with medical-grade ECG devices at the feature level in their preliminary study. Schuurmans et al. [14] assessed the validity of *Empatica E4* devices under baseline conditions again at the feature level. After these studies compared wearable devices with medical-grade ones under baseline conditions, researchers realized that the validation should be extended to more challenging conditions such as mental or physical stress. Konstantinou et al. [26] compared wearable devices with stationary devices under mental stress conditions at the feature level. They used Microsoft Band 2 as the wearable and showed that the features' distances with the reference device increase under mental stress. After considerable effort at the feature level, researchers started preliminary studies at the application level. Ollander et al. [27] compared the *Empatica E4* device with a stationary sensor under mental stress with seven participants. They stated that although the stationary sensors are very close to the reference devices at feature levels when compared to wearable devices, the stress discrimination power at the application level is maintained with wearable sensors. Furthermore, in another study [28], the authors trained machine learning models to compare the Biopac MP150 (laboratory sensor) and Empatica E4 (wearable sensor) for recognizing emotions. They reported that the accuracy of emotion recognition with data collected by wearable devices was very similar to when data was collected by laboratory equipment. These comparison studies promoted wearable devices and led to more research on emotion and stress recognition using wearables for data collection [29, 30, 31].

Physical activity is another condition that wearable sensors should be validated at feature and application levels. Pinheiro et al. [32] found an average correlation of 0.82 between features extracted from wearable and medical-grade sensors. Even in physically active conditions, they reported a 0.68 correlation which shows a relatively strong relation. Furthermore, they computed a 0.88 correlation between these signals in the resting conditions. They concluded that the presented results show the potential of the PPG sensor as an alternative for HRV analysis in healthy subjects, even in physically active situations. As seen in the literature (see Table 1), although there are preliminary studies with a limited number of devices and subjects comparing wearables with stationary devices, there is no study comparing the wearable devices under mental stress and physical activity at both feature and application levels.

Our study is not only the first one to compare wearables under mental stress and physical activity at both levels, but we also used the most comprehensive and diverse set of wearables in the literature for comparison. We also employed a state-of-the-art explainable AI (XAI) method to investigate and demonstrate

Table 1
List of comparison studies conducted using wearable heart monitoring sensors

Article	Comparison type		Monitoring type			Sensors		
	Application-level	Feature-level	Mental stress	Physical activity	Number of participants	Multi-modal	Number of sensors	Placement
[27]	✓	✗	✓	✗	7	✓	3	Wrist
[23]	✗	✓	✗	✗	2	✗	2	Chest, Finger
[24]	✗	✓	✗	✗	10	✗	2	Chest, Finger
[32]	✗	✓	✗	✓	35	✗	2	Chest, Finger
[28]	✓	✓	✓(Arousal)	✗	19	✓	2	Chest, Finger
[15]	✗	✓	✗	✗	49	✗	2	Finger, Wrist
[26]	✗	✓	✓	✗	43	✓	2	Wrist
[14]	✗	✓	✗	✗	15	✗	2	Chest, Wrist
[25]	✗	✓	✗	✗	19	✗	2	Wrist
[20]	✗	✓	✗	✗	32	✗	5	Chest, Wrist
Current study	✓	✓	✓	✓	32	✓	7	Chest, Wrist, Finger

Check mark symbols (✓) are used to indicate "Yes", and X mark symbols (✗) are used to indicate "No".

the importance of features and data preprocessing techniques and their impact on the output of various classification algorithms.

3. Methodology

The primary purpose of this study is to compare the performance of different wearable devices in the application of mental and physical stress diagnosis. This comparison was performed in terms of accuracy and sensitivity of diagnosis. In addition, several factors, such as differences in the impact of features, have been examined and compared, which will be explained in detail in future chapters.

Overall, We recruited 32 healthy participants (10 Females and 22 Males, age=28.4±5.98 years, BMI=25.61±6.49) for data collection through advertisement within our university campus. Participants were instructed to maintain a regular sleep schedule the night before [33], and refrain from eating food and consuming caffeinated drinks two hours prior to the study [34, 35]. The data acquisition procedure took place in a soundproof room without any auditory and visual distractions, facilitated by the first and third authors during participants' single visit to our lab. Each participant's session took 70 minutes and was divided into four parts, as detailed below.

3.1. Part 1: Study Introduction and Participants' Consent

The first part of the study consisted of researchers introducing the study to the participants. Each participant was given a participant information sheet that detailed all the information regarding the data collection procedure, i.e., what data will be collected and how it will be used. Moreover, participants were told that their participation was voluntary, and they were free to withdraw during and after a week of the study without giving any reason. Participants were notified that if they chose to leave the study during this time period, all of their data would be destroyed. If they agreed to participate, their data would be anonymized and would not link to their names or any other details that could identify them. Participants were encouraged to ask questions at this stage if they did not understand anything. Once participants understood the study details, they signed a consent form followed by a short demographics questionnaire. All participants underwent exactly the same sessions during the data collection sessions, and each visited the laboratory only once. The procedure of the methodology used in this study complies with the 1964 Declaration of Helsinki [36] and is approved by the Institutional Review Board for Research with Human Subjects of Boğaziçi University with the approval number 2018/16.

Table 2
Heart monitoring sensors used in this study, 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
[37, 38, 39, 40]	Empatica E4	PPG, EDA, ACC, IR Thermopile	64 Hz	Wrist	Bluetooth	✓	✓	✗
[41, 42, 43]	Samsung Gear S2	PPG, ACC, Barometer, Gyro	100 Hz	Wrist	Bluetooth	✓	✗	Display
[44, 45, 46, 47]	Firstbeat Bodyguard 2	ECG, ACC	1000 Hz	Chest	USB	✗	✗	✗
[48, 49, 50, 51]	BITalino (r)evolution	ECG, EEG, EDA, EMG, ACC	1000 Hz	Chest	Bluetooth	✓	✗	Vibration, Buzzer, Led
[52, 53, 54, 55, 56]	Polar H10	ECG	130 Hz	Chest	Bluetooth	✓	✓	✗
[57, 58, 59, 60]	Zephyr HxM	ECG	250 Hz	Chest	Bluetooth	✓	✗	✗
[61, 62, 63, 64]	CorSense	PPG	500 Hz	Finger	Bluetooth	✓	✓	✗

Check mark symbols (✓) are used to indicate "Yes", and X mark symbols (✗) are used to indicate "No".

3.2. Part 2: Sensor Placement

The second part of the study consisted of sensor placement. Two of the authors helped participants place all the sensors as shown in Figure 1. We used the standard placement procedure recommended by the manufacturer for data recording while ensuring that it was comfortable for participants. Seven different off-the-shelf heart rate monitoring wearable devices were utilized in this study: *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 2. 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. Further information on the sensors and data acquisition are detailed in section 3.5.

3.3. Part 3: Mental Stressor

Once all the sensors were placed correctly and ready for recording, we started recording the baseline stress level measurement data for 10 minutes. In order to achieve a precise and accurate interpretation of psychophysiological phenomena and of vagal tone, and based on the guidelines in the literature [65], participants were instructed to avoid any large movements while sitting comfortably in an upright posture. After the baseline session, participants were then asked to perform stressor tasks for ten minutes, followed by five minutes of resting, cycling, and finally, the last resting phase, as shown in Figure 2. In phase one of the stressor task, participants were taught how to complete the Stroop Color Test on a tablet. The Stroop Color Test is a color and word neuropsychological test widely utilized in the literature and clinical purposes [66].

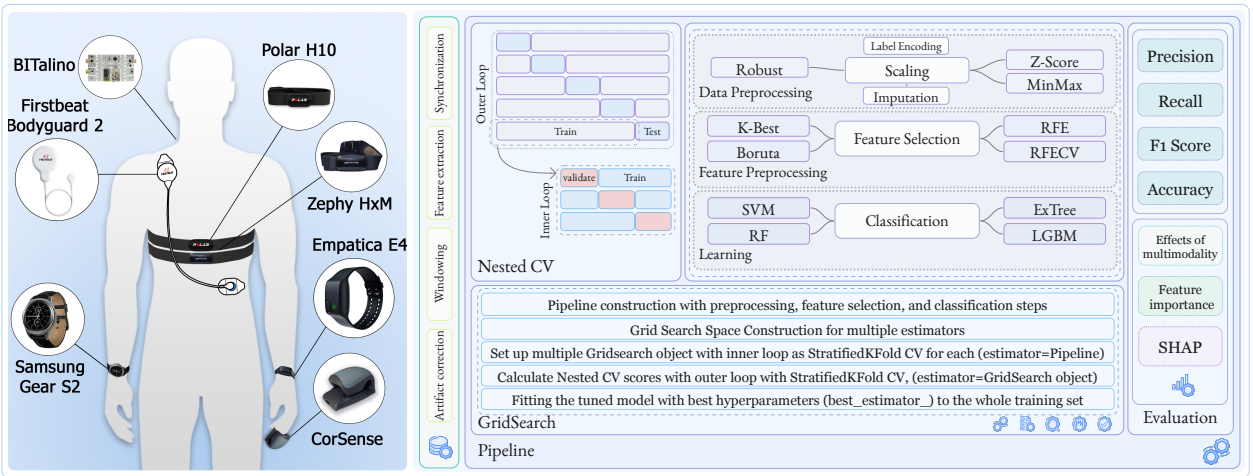


Figure 1: A brief outline of our system architecture demonstrating several steps. Starting from the left with sensor placement

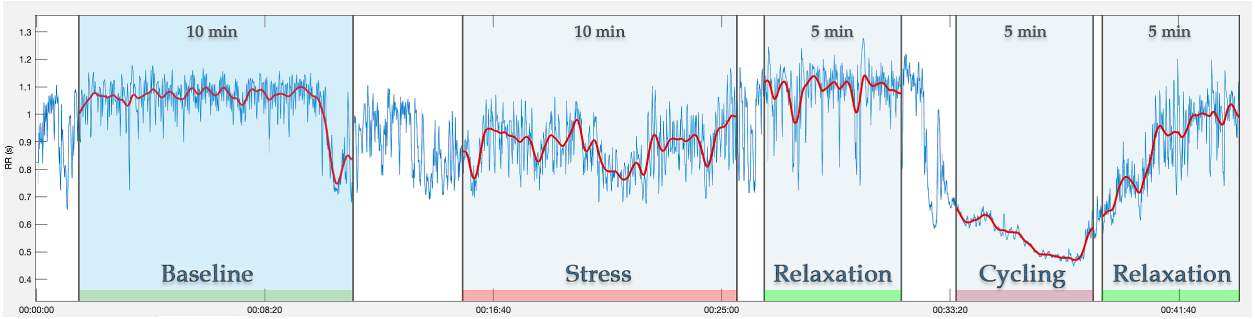


Figure 2: Sample RR intervals with ten and five minute windows collected from a single subject. It should be noted that the gaps in between the activity windows correspond to the time between the activities

The Stroop Color and Word Test (SCWT) is based on the principle that humans can read words more quickly than they can identify and recognize colors. During the Stroop Color Test, color names (e.g., green, red, or blue) appear in different colors (e.g., the word "blue" in green, rather than in blue). The subject is asked to identify the color words displayed in an inconsistent color (for instance, 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 [67, 66]. Participants were asked to carry out this test for five minutes and also were told that their test results would be compared to those of other participants. It must be noted that SCWT has been used at different degrees of difficulty in some studies. For instance, de Arriba-Pérez et al. used three difficulty levels, where colors and words match, colors and words still match but with different randomized order of appearance, and colors and words do not match, are levels one, two, and three respectively [68]. Compared to de Arriba-Pérez et al.'s work [68], the level of SCWT difficulty in our study's stress session was level three, and a combination of levels one and two was only used during a short training for each participant before the start of the stress session.

As a follow-up to the Stroop test, the second phase of the stressor task consisted of arithmetic questions, a component of the widely known Trier Social Stress Test (TSST) [69]. While one of the authors pretended to take notes of the correct and incorrect responses, participants were asked to count backward with varying differences for five minutes (for example, counting backward from a particular three-digit number in steps of 23). At the end of this session, participants were asked to rest for five minutes. Both TSST and Stroop color test, which are highly popular and widely accepted in the scientific community, are very effective methods to induce psychological stress [16]. We utilized these two different stressors for five minutes each in order to prevent the possibility of habituation and fatigue. When the physiological signals are examined (see Figure 2), it could be seen that both methods successfully created stress reactions in participants.

3.4. Part 4: Physical Stressor

The third part of the study consisted of a physical stressor where participants were asked to perform cycling on an indoor exercise bike.

The reason for choosing cycling as a physical activity in our study was 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 sympathetic and parasympathetic components of the autonomic nervous system. When an individual engages in exercise and physical activity, it affects their HRV, as physical demands are met by both components [70]. HRV has been used in the literature for measuring physical stress [71, 72, 73] with cycling as the exercise method or even psychological stress during the cycling [74].

The Cycling activity in our study 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.

3.5. Sensors

Firstbeat Bodyguard 2 serves as one of the ECG devices in our experiment. This lightweight wearable sensor for measuring cardiac signals and R-R intervals is validated in [75] and used in many studies in recent years. Once the *Firstbeat Bodyguard 2* is connected to the skin, it begins recording 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, an ECG chest strap that can provide an accurate heart rate measurement at a frequency of 130 Hz, is the second ECG wearable employed in this study. Using the *Polar H10*, the RR data can be recorded in real-time on a smartphone and saved in cloud storage as well. Additionally, it is capable of transmitting its RR data to the *Polar V800* sports watch and heart rate activity monitor. The third device we used in this study was a board kit produced by PLUX Wireless Biosignals. The *BITalino (r)evolution* board kit includes multiple types of sensors and actuators and measures the ECG at a speed of 1000 Hz. Acquisition of the ECG signal in BITalino kit is performed live via Bluetooth connection with a computer. The fourth ECG device employed in this study is *Zephyr HxM*. 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. All of the ECG devices mentioned above, except the BITalino kit, provide RR values processed inside the device instead of raw ECG signals. *Empatica E4*, *Samsung Gear S2*, and *CorSense* are the PPG devices utilized in this study. *Empatica E4* has a charging time of fewer than two hours and a battery life of 32+ hours. Additionally, it can store 60 hours of data, and as well as a USB connection, it can perform real-time data transfers through Bluetooth. The *Empatica E4* is exclusively designed for research purposes and is equipped with additional sensors like EDA, body temperature, and accelerometer. *Empatica E4* uses a 64 Hz sampling rate to record the raw blood volume pulse (BVP) signal with its PPG sensor [76]. The inter-beat interval values are calculated using the diastolic points of the blood volume pulse, which correspond to the local minima of the BVP signals. The second PPG device utilized in this study is *Samsung Gear S2*. It is a smartwatch capable of generating IBI values from the signals captured with its PPG sensor. Gear 2's battery provides up to three hours of continuous use when the PPG sensor is activated for continuous recording. The third PPG sensor used in this study is *CorSense*, an HRV monitor and live biofeedback training device designed to detect heart rate signals from the fingertip. It measures heart rate variability through a PPG sensor at 500 Hz. Two of the authors present at the lab instructed participants on how to wear the device, and participants were also advised not to make sudden movements during data collection. Details mentioned above were only parts of the on-the-paper device specifications advertised by the manufacturer. We will share our hands-on experiences gained with all devices during the data recording sessions in the following chapters.

4. Data Analysis and Results

This section outlines the data and various operations we conducted to obtain the most accurate and reliable results. Following the comprehensive description of the various devices provided in the previous section, here in this section, we will provide a brief description of the practical experiences that we gained during data collection with these devices. The raw data will be described, and the operations carried out for data pre-processing. Next, we will discuss the methods for pre-processing the data before classification. Next, we will discuss the methods and algorithms utilized to select the features, construct the pipelines, and find the best hyperparameters. Last but not least, we will discuss the results and explain them comprehensively from a statistical and explainable artificial intelligence perspective. Several steps of our system architecture

are illustrated in Figure 1, starting with the preprocessing of signals and raw data on the left, it progresses with the ML pipeline and concludes with XAI and performance metrics.

4.1. Data Labeling

Stress levels reported by an individual can be under the influence of factors such as social-desirability bias, and high subjectivity of stress, resulting in exaggerated or understated reports [39, 77]. It has been shown that even different genders may report their emotional states differently when faced with the same situation [78]. Since the stressors in this study (TSST and SCWT) were induced precisely in the same way to all participants, we decided to use the context information (known activity type) as the data labels to prevent the probability of bias in subjective reports. Our data are accurately labeled using known context information. This is due to the fact that data collection was carried on in the laboratory environment and two of the authors were always present throughout all data acquisition sessions and kept detailed notes on the exact start and end times of each session and activity type.

The experiment was designed based on different levels of mental and physical stress. We included baseline ("low stress") and TSST+SCWT ("high stress") states. Furthermore, because the majority of the studies only differentiate between stress and baseline states, which is not representative of the vast array of possible nuances that define the continuum from baseline to the highest stress levels, the "medium stress" level was also added to the experiment design. We included the recovery (relaxation) state for the "medium stress" level. This state begins right after the "high stress" level, and it is expected to return participants to their baseline ("low stress") levels using relaxation and breathing-based meditation. Right after a ten-minute intense stress session, the recovery state that follows it would definitely consist of lower stress levels than the preceding "high stress" state and higher than the participant's baseline ("low stress") measure. We also included physical activity ("physical stress") because the physiological response to physical activity resembles the physiological stress response. The states of relaxation ("medium stress") and physical activity ("physical stress") are added because there must be states similar to, or likely to be confused with, "high stress" while designing effective ML models which can classify four stress classes.

4.2. Data Preprocessing

The standard duration of short-term HRV recordings is five minutes as recommended by the North American Society for Pacing and Electrophysiology and the task force of the European Society of Cardiology [79]. In this study, we carried out the HRV analysis on basic and short-term durations, ten and five minutes, respectively. Based on the rationale explained in the previous paragraph, and using timestamps collected during the data acquisition session, we divided the raw heart rate data into five consecutive sessions, i.e., Baseline, Stress, Recovery 1, Cycling, and Recovery 2, as shown in Figure 2. This was followed by signal synchronization. To carry out the preprocessing step of the analysis, "*Kubios HRV*" version 3.4.1, a scientifically validated HRV analysis software, was utilized [80]. *Kubios HRV* is a robust HRV analysis tool used by researchers to study cardiac data and assess stress's effect on human well-being.

The value of the data becomes even more apparent in studies that focus on human health data and where research findings rely solely on the analysis of this data. It takes a great deal of effort to adequately prepare and safeguard the data that has been collected with great difficulty and much time spent. One of the most crucial requirements in a biosignal acquisition device is its ability to facilitate access to and ensure the integrity and availability of the recorded data. Different devices store their data with different mechanisms and across varied locations. Given that we are conducting a comparative type of research, remarking details concerning the differences in data storage and retrieval mechanisms between the various biosensor devices as well may be beneficial for the end-users. In different applications, the level of expectations is quite different. For instance, when collecting data from a large number of users for research purposes or even for personal use on a daily basis. The degree of complexity and time required to access data can be an essential criterion in selecting one type of device over another. The fact is that for the end-users, in addition to the accuracy of measurement and

data quality, minor differences such as cloud storage options can also make a big difference in their decision to choose a specific type of device.

Data is simply stored in the device's internal storage in three of the devices we utilized. These devices are *Firstbeat Bodyguard 2*, *Empatica E4*, and *Samsung Gear S2*. Data cannot be saved locally and accessed after the experiment in the rest of the devices. For accessing data recorded by *BITalino (r)evolution*, *Polar H10*, *CorSense*, and *Zephyr HxM*, a direct wireless connection needed to be established with another device, such as a computer or smartphone.

In order to maintain the integrity of the test conditions and data collection, we could not allow any interruptions, modifications, or pauses during the recordings. Each session must have been continued until the end without any intervention, immediately after it started. Therefore, in the event of a potential drop in the wireless connection, we were unable to pause the session and resolve the issue, and the data from the disconnected device must be considered as lost from that particular participant. However, all devices retained stable connections, and for all 32 participants, wireless connections of all devices were maintained intact. Working with devices with internal storage was a lot easier. Nevertheless, there were a few exceptions. For example, *Firstbeat Bodyguard 2* could not detect individual users in a situation where multiple users used the device almost immediately after each other. In such a case, after collecting the data, we had to manually separate the data of different users from a single file for several consecutive sessions using timestamps. The Samsung smartwatch does not have access to IBI data by default, so we used an application developed by Can et al. [81] to store the IBI data in the device's internal storage. *Empatica E4* is equipped with the best storage method, in which data is both saved internally and uploaded to a free cloud space provided by the manufacturer. Due to the fact that in studies like this, the data is genuinely invaluable, its loss can be considered as an irreparable waste of time and resources. Therefore, we made every effort to store the data as efficiently as possible and store it appropriately.

4.2.1. RR detection

One of the essential steps in analyzing cardiac data is to access the R-R peaks. The stored data from the ECG and PPG sensors come in a variety of formats. Some devices, such as the *BITalino (r)evolution*, only store large volumes of raw data without any internal preprocessing. Thus, further processing is required to extract R-R intervals data from the ECG data collected from such a device. The same is true in *Empatica E4*, in which we must process BVP data to obtain IBI (RR). Processing takes place inside the device in all other devices, and the ready-to-use data is then delivered to the user. Although the latter takes much less time to preprocess in the later stages, we are restricted to using the R-R intervals data generated by the device. In contrast, we can use more advanced and more efficient algorithms to generate R-R data with devices that provide raw files. As described above, in output files created by devices that only provide raw data, no R-R intervals data is present. Therefore, we need to convert these raw data into R-R intervals using another method. For instance, the premium version of *Kubios HRV* version 3.4.1 was used to perform R-R extraction from raw files produced by *BITalino (r)evolution*. As a powerful tool frequently used by the research community, *Kubios HRV* is considered as one of the most widely used tools for analyzing heart health data and measuring stress's impact on health, and well-being [80]. The other devices, in which the RR intervals data were computed directly by the device itself, were also preprocessed by *Kubios HRV*, including feature extraction and artifact correction.

4.2.2. Artifact removal

One of the most common problems in using wearable devices in daily life is the susceptibility of these devices to noise. In the HRV related studies, users' involuntary fast movements and environmental factors may contribute to these noises, and different types of sensors can be affected differently depending on their attachment locations and sensor types. These adverse effects and noise that negatively affect data quality and clearance are known as artifacts. In order to diminish the chances of adverse effects that these artifacts

can induce in the HRV data and stress assessment, the Task-force of the European society of cardiology recommends that all artifacts in the R-R time series data be either corrected or removed [79].

The artifact correction algorithm available in *Kubios HRV*, which we applied to the data from our seven wearables, is a validated novel method that can detect different types of artifacts with a sensitivity over 96.96% [82]. It is based on time-varying thresholds calculated from the series of succeeding RR-interval variations paired with a beat classification method. In a previous study, we discovered that using appropriate noise-canceling algorithms with proper thresholds led to promising results [20]. When applied to data from PPG devices, it resulted in gaining higher correlation and agreement levels with data recorded from ECG devices. The extent to which individual sensors are influenced by ambient noise alters the nature of their data. Therefore, a single solution cannot be applied to all devices, and different noise reduction thresholds were applied to the data coming from different devices. As expected, ECG devices produced the smallest amount of artifacts of all. Some of them, such as *Firstbeat Bodyguard 2*, also execute internal noise-reduction methods on their data and present both raw and corrected data in the output file. To avoid discrepancies in the type of algorithms employed and to ensure the integrity of the operations conducted on the data, we used the raw output in this study and performed noise reduction by *Kubios HRV* when needed.

The artifacts' presence was substantially higher in smart and non-smart wristbands, all equipped with PPG sensors. The amount of artifact in the *CorSense* finger sensor was lower than in other PPG devices. Perhaps this is due to the fact that it is designed to be used for at rest measurements. In its user manual and even during the connection establishment with the smartphone, warning messages are presented to the user to minimize hand and finger movements. Hence, we used *CorSense* only in the first three sessions, and participants were asked to remove the device from their fingers at the start of the cycle session.

4.2.3. Synchronization and windowing

When comparing signals of the same nature recorded from different sources, signal synchronization is essential in order to ensure no time difference exists between the recordings from the reference device and those from the comparison device. All sessions data from each one of the devices were synchronized precisely to achieve an accurate and fair comparison. Data synchronization was performed both automatically and manually during data preprocessing to ensure absolute alignment. We determined the time shift between the signal from the devices using cross-correlation, which is a widely used method in similar studies [83, 84, 13]. In this method, the time-point when the maximum of cross-correlation is obtained corresponds to when the signals are best synchronized. Besides, *Kubios HRV* 's data browser environment, as shown in Figure 2, was also used for visual inspection and manual alignment of the raw peaks. Similar studies have employed this method as well [85, 18], and we found out that in our analysis, this method was as accurate as cross-correlation. It should be noted that apart from the accurate time stamp provided by each device, two of the authors were present throughout all data acquisition sessions and kept detailed notes on the exact start and end times of each session.

HRV features were calculated with a window width of 300 seconds and a grid interval of 60 seconds for the moving window. The selection of 300 seconds windows was due to the fact that calculating HRV features such as the root mean squared difference of successive R-R intervals (RMSSD) requires a minimum window of five minutes [79, 86].

4.3. 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 [87]. 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 in order to bring all features to the same level and make sure 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 process of standardization of the data.

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 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 kinds of scaling algorithms, each employing its own method for estimating the parameters for shifting and scaling the data. In this study, we tested a number of 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.

4.4. 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 that are far from the truth.

4.4.1. Feature selection

Feature selection is a crucial concept that can significantly impact a model's performance by removing irrelevant features. In order to lessen the complexity and reduce the time required for the execution of computations, which has been greatly increased due to the utilization of Nested cross-validation, feature

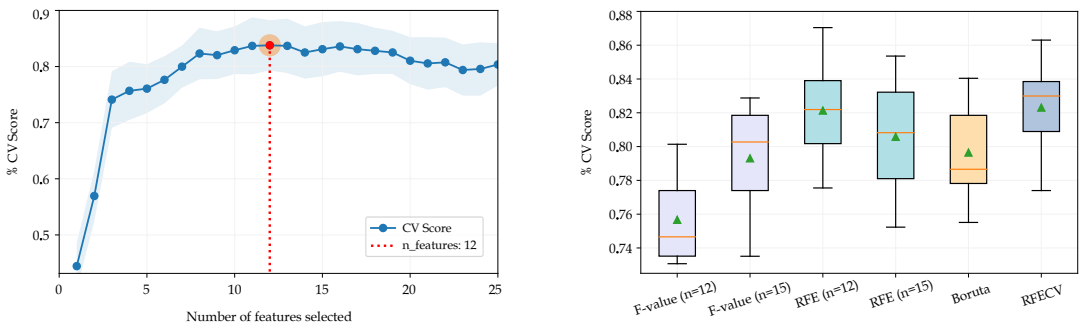
Table 3
List of the features selected by Recursive Feature Elimination with cross-validation (RFECV)*

Feature name	Feature type			Description
	Time-domain	Frequency-domain	Nonlinear	
Mean RR *(ms)*	✓			The mean of RR intervals
STD RR *(ms)*	✓			Standard deviation of RR intervals
RMSSD *(ms)*	✓			Square root of the mean squared differences between successive RR intervals
HRVti	✓			The integral of the RRI histogram divided by the height of the histogram
TINN *(ms)*	✓			Baseline width of the RR interval histogram
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
SD2/SD1			✓	Ratio between SD2 and SD1
ApEn			✓	Approximate Entropy
SampEn			✓	Sample Entropy
HR Max - HR Min	✓			Difference of the Minimum and Maximum HR

Features in this table are not ordered by importance.

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 in model accuracy. To make a fair comparison, 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 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 3a, and Table 3.

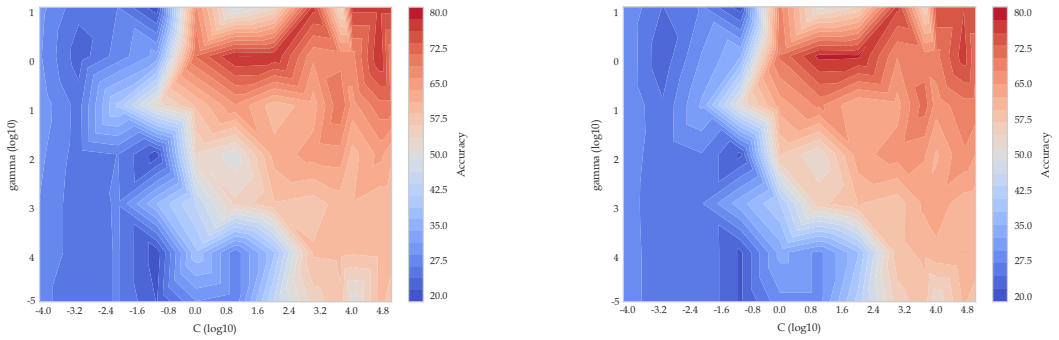
Ideally, it is better to select the features inside the ML pipeline. However, running RFECV inside the pipeline could lead to the selection of very distinct sets of features for every model. Therefore, in order to keep the comparison conditions equal in terms of the number and the type of features selected for all devices,



(a) Number of features selected by Recursive Feature Elimination with Cross Validation

(b) Effects of different methods and the number of optimal features selected by each method

Figure 3: The optimal method and number for feature selection



(a) Polar H10 (b) Empatica E4

Figure 4: Hyperparameter optimization in SVM for two different devices

the selection of 12 features was performed outside the pipeline by Recursive Feature Elimination (RFE). The decision for choosing 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 3b.

4.4.2. 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 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 for achieving quick results and much higher speeds of the operations. Figure 4 illustrates the effects of different combinations of hyperparameters on the SVM classifier's output accuracy for two different devices. 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.

4.4.3. 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, in the process of tuning the hyperparameters and evaluating model performance, cross-validation uses the same set of data. This can result in data leakage and lead to unintended overfitting of the model, and overly optimistic biased results [88]. 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. Data leakage problem and the optimistic bias caused by it can be avoided by nesting the hyperparameter optimization procedure beneath the model selection [89]. This procedure is known as nested cross-validation. A nested cross-validation strategy allows for a more robust

and generalized assessment of model performance [90]. 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 generalization error. The inner loop is used for *GridSearchCV* object and the *cross_val_score* object uses the outer loops.

4.5. 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 second 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, and the random forest output is the class chosen by the majority of trees [91, 92]. 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 that 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 [93]. 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 [94]. 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 speed and efficiency and 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 [94, 95]. 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.

The primary purpose of this study is the assessment and comparison of 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.

4.5.1. Reproducible splits

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

Table 4
Classification results using four algorithms for all seven wearable devices

		SVM			Random Forest			ExtraTree			LightGBM		
Device	**Session**	**Precision**	**Recall**	**F_1-Score**	**Precision**	**Recall**	**F_1-Score**	**Precision**	**Recall**	**F_1-Score**	**Precision**	**Recall**	**F_1-Score**
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%		

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).

4.6. 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 4 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 [96]. However, as already described in 4.4.3, 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 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 [96]. 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. 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 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 4 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 4.5, 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 [20] maintained their superiority in stress classification applications as well. In the ExtraTree classifier, the average test accuracy

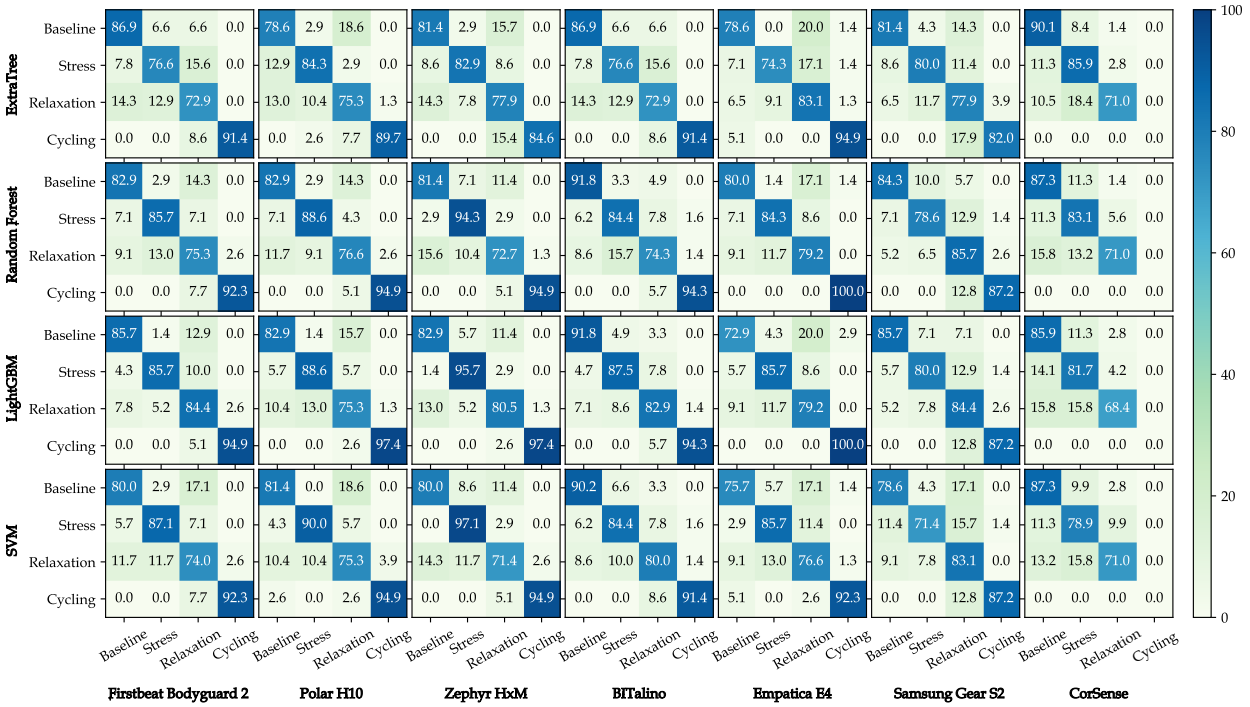
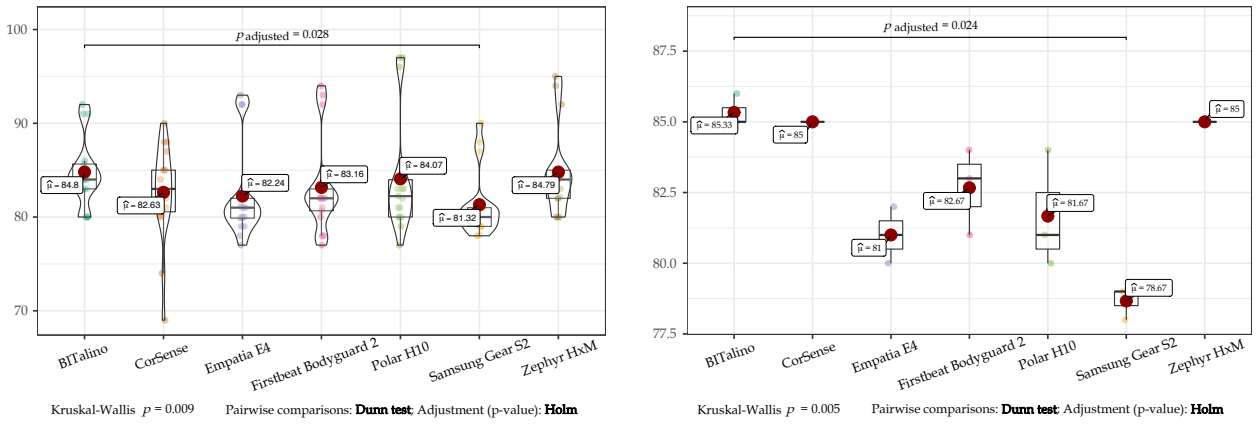


Figure 5: Normalized Confusion Matrix computed using four classifiers for all devices



(a) All sessions - LightGBM classifier

(b) Stress - LightGBM classifier

Figure 6: Kruskal-Wallis comparison followed by Dunn's test for all devices with the LightGBM classifier

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 4 provide a brief overview of the overall differences between our seven devices. Moreover, the normalized confusion matrix for all models is depicted in Figure 5. However, we need to take further steps to make a valid judgment 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 they did not meet the assumptions of normality, we decided to utilize the Kruskal-Wallis test, a non-parametric equivalent of one-way ANOVA [97, 98]. As seen in Figure 6a, 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 [99]. For the results obtained with LightGBM for all classes, posthoc analysis showed significant differences only between *BITalino (r)evolution* and *Samsung Gear S2* wearables ($p = 0.028$). 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 best results (*BITalino (r)evolution*) and the PPG wearable with the lowest results (*Samsung Gear S2*), and there exists no statistically significant difference between other devices. Repeating the same statistical approach to analyze the results obtained from the ExtraTree classifier results in more conservative results. 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*).

4.6.1. Effects of Multimodality

Wearable devices in this study showed great potential for producing reliable results for stress detection and classification, 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.

Although the idea of using Electrodermal Activity (EDA) in psychological research has been around for a long time [100], it is still among the top biosensing measures employed in the subject area of affective computing using ubiquitous and wearable devices [12, 101]. Electrodermal activity (EDA) is an umbrella term for Galvanic skin response (GSR). 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. Therefore, skin conductance acts as a measure of sympathetic and emotional responses. 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 using NeuroKit2, which is a Python toolbox designed for neurophysiological biosignal processing [102]. As seen in Figure 7, Tonic and Phasic components of the EDA signals and their corresponding features were extracted using NeuroKit2, and a closer look at the EDA data from five subjects is demonstrated at the bottom of the

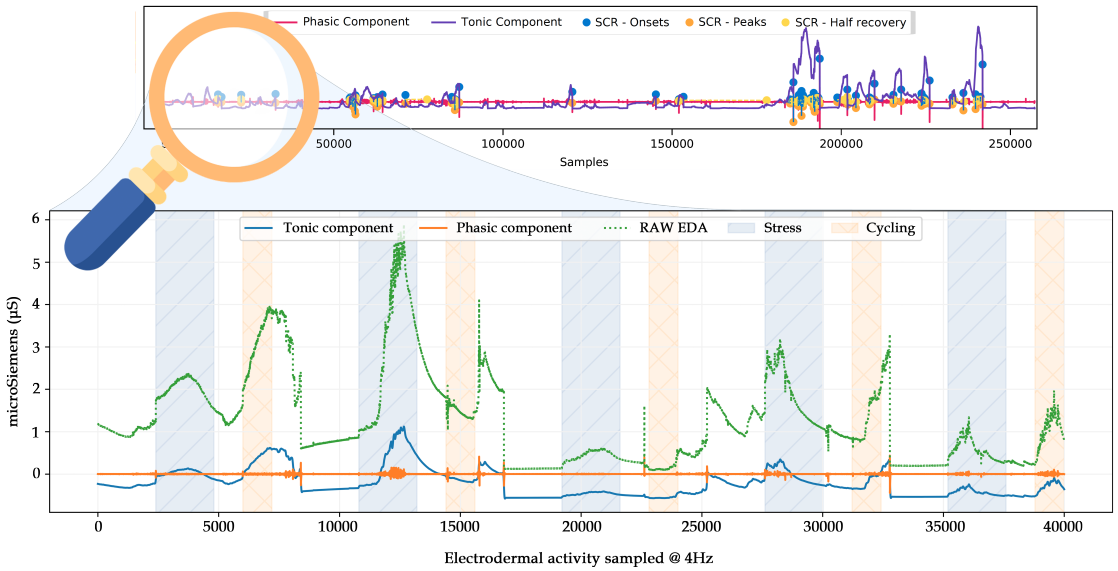


Figure 7: Sample of Electrodermal Activity for five participants

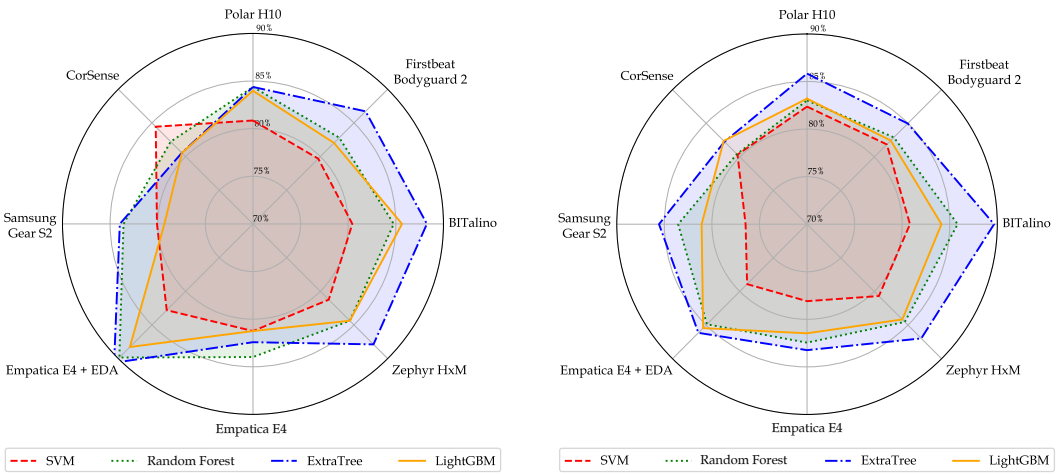
Table 5
Classification results using four algorithms for the Empatica E4 with and without EDA data

		SVM			Random Forest			ExtraTree			LightGBM		
Device		**Precision**	**Recall**	**F_1-Score**	**Precision**	**Recall**	**F_1-Score**	**Precision**	**Recall**	**F_1-Score**	**Precisio**	**Recall**	**F_1-Score**
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 + 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%		

same figure. Based on the results reported in Table 5, and displayed in Figure 8, 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 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.

4.7. Comparison with the literature

Recently, several studies have been published to explore the potential of commercial wearables in detecting different types of stress. We listed the prominent ones in Table 6. As seen in this table, 2-class stress detection results (stress vs. baseline) are above 90% in laboratory environments. Our study obtained a maximum of 90.6% accuracy in four class classification (stress, baseline, relaxation, and physical activity), which is clearly a more difficult task. However, different studies in the literature have been applying various methods to the unique datasets collected for their study. Therefore, unless the same data sets are used, one



(a) Accuracy performance comparison (b) F_1-Score performance comparison

Figure 8: Comparison of Test Accuracy and F_1-Score results using four classification algorithms on data from seven devices

Table 6
Recent Stress-related studies using physiological signals.

Article	Biosignal	Algorithm	Number of Classes	Accuracy*	Environment	
					Daily Life	Laboratory
[103] (2021)	HRV	KNN, SVM, MLP, RF, XGB	2 (Baseline, Stress)	76	✓	
[104] (2021)	HRV	KNN, SVM, MLP, RF, GB	2 (Stress, no Stress)	80	✓	
[31] (2021)	EDA	SVM	4 (Baseline, and three Stress stages)	75		✓
[105] (2020)	HRV, EDA, EMG, ACC, RESP	SVM, KNN, Adaboost, FNN	3 (Amusement, Baseline, Stress)	84.32		✓
[106] (2020)	EDA, HRV	RF	2 (Stress, no Stress)	92		✓
[107] (2019)	HRV, EDA	SVM, RF	2 (Baseline, Stress)	77		✓
[29] (2019)	HRV, EDA, ACC, ST	LDA, QDA, RF	2 (Baseline, Stress)	87.4		✓
[30] (2018)	HRV, EDA and ST	SVM, C4.5, kNN, RF, NaiveBayes	2 (Baseline, Stress)	99.85		✓
[81] (2019)	HRV, EDA, ACC	MLP, RF, KNN, SVM	3 (Baseline, Cognitive Load, Stress)	92.15	✓	
[108] (2017)	HRV, EDA, RESP, SPO2	SVM, KNN	2 (Baseline, Stress)	95.8	✓	
[109] (2016)	EEG	SVM	4 (Neutral, Low, Medium, High Stress)	89	✓	
Current study	HRV, EDA	ExtraTree, LightGBM, RF, SVM	4 (Baseline, Stress, Relaxation, Cycling)	90.6		✓

Check mark symbols (✓) are used to indicate "Yes".

KNN: k-nearest neighbors, SVM: Support vector machine, MLP: Multilayer perceptron, RF: Random forest, XGB: XGBoost, GB: Gradient boosting, FNN: Feedforward Neural Network, ACC: Accelerometer, EMG: Electromyography, RESP: Respiration, EEG: Electroencephalogram

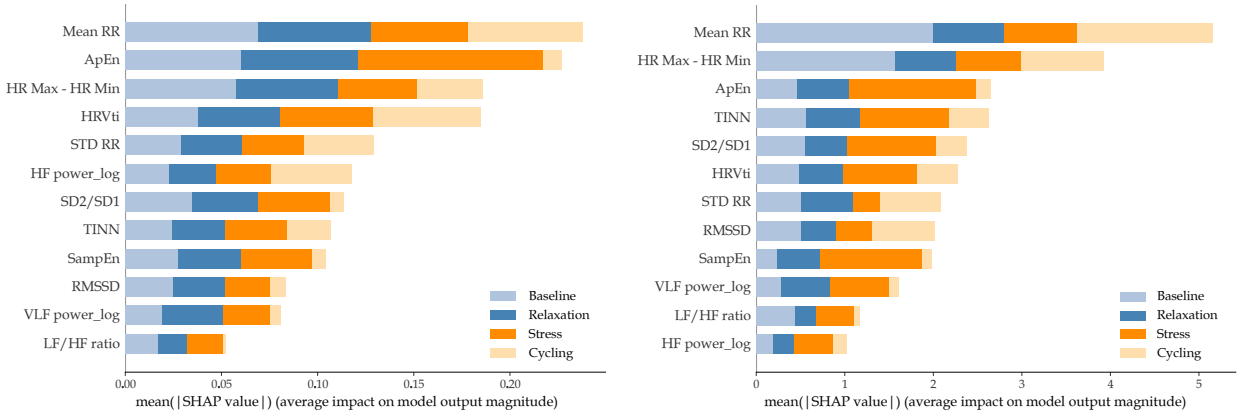
* Best achieved Accuracy

could not infer the success of a technique over the others. These results were presented to provide the reader with a sense of the performance of the stress detection studies in different environments. Having said that, it could be stated that the proposed results are aligned with the best-reported results in the literature.

4.8. Model explainability

In many applications, understanding why a model reaches a particular prediction is just as essential as its accuracy. Nowadays, it is often possible to achieve high accuracies with very complex models. However, the model behavior and identifying the factors involved in the outcome becomes very hard to interpret [110]. 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. Despite the fact that ML models have made it considerably easier to predict the feature health conditions of an individual, they still fall short in interpretability [111]. Identifying and interpreting which features contribute most to a particular prediction in different models can be very useful, especially 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 [112]. An essential aspect of investigating the models' explainability in the context of this study is to determine whether the 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 designed in the area of Explainable Artificial Intelligence (XAI) [113, 114, 110]. These tools have evolved for model interpretation and demystifying black-box models over the last few years and are becoming more popular each day.

In this study, we utilize SHapley Additive exPlanations (SHAP) for explaining our models. SHAP is an open-source game-theoretic approach to explain the results of ML models based on game theory. It can probably be considered as 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 [110]. Application of SHAP analysis to our data and our 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 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



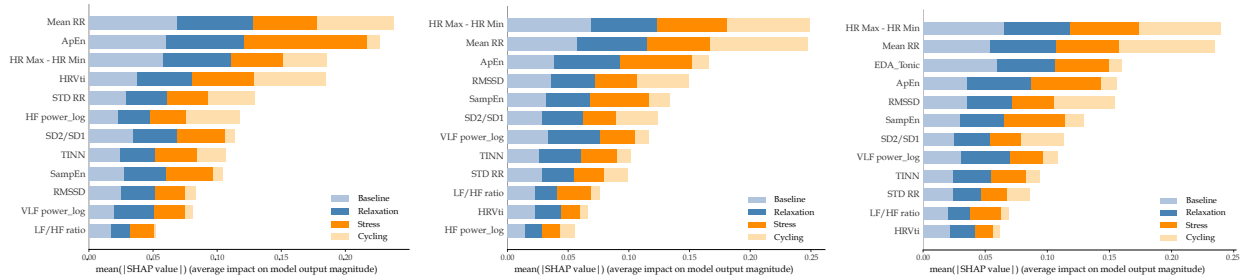
(a) ExtraTree Classifier

(b) LightGBM Classifier

Figure 9: Feature influences with SHAP on all classes, with two models using the Firstbeat Bodyguard 2 wearable device

and extend it to its application in this study [115]. 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. By this means, 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 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 of the top-performing models, ExtraTree, and LightGBM, respectively. Data for these two models came from the *Firstbeat Bodyguard 2* device. 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 the second place with the ExtraTree classifier, it is in the third place using LightGBM, and basically, 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 classifier, Mean RR influences the prediction of each class in almost the same magnitude, in LightGBM, this influence is doubled for the Baseline and Cycling sessions. ApEn has the most significant influence on predicting the Stress class in both models, and the frequency-domain feature LF/HF ratio has little to no impact on predicting the physical stress (Cycling) session. As a final example for Figure 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



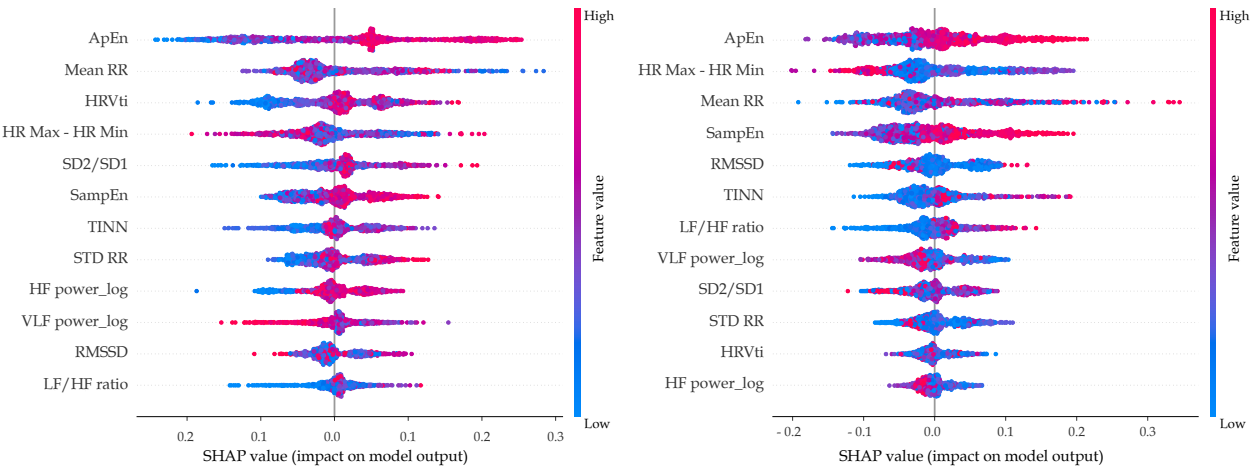
(a) Firstbeat Bodyguard 2 - ExtraTree (b) Empatica E4 - ExtraTree (c) Empatica E4 + EDA - ExtraTree

Figure 10: Feature influences with SHAP on all classes, with ExtraTree classifier using the ECG, PPG, and PPG + EDA wearable devices

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, 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.

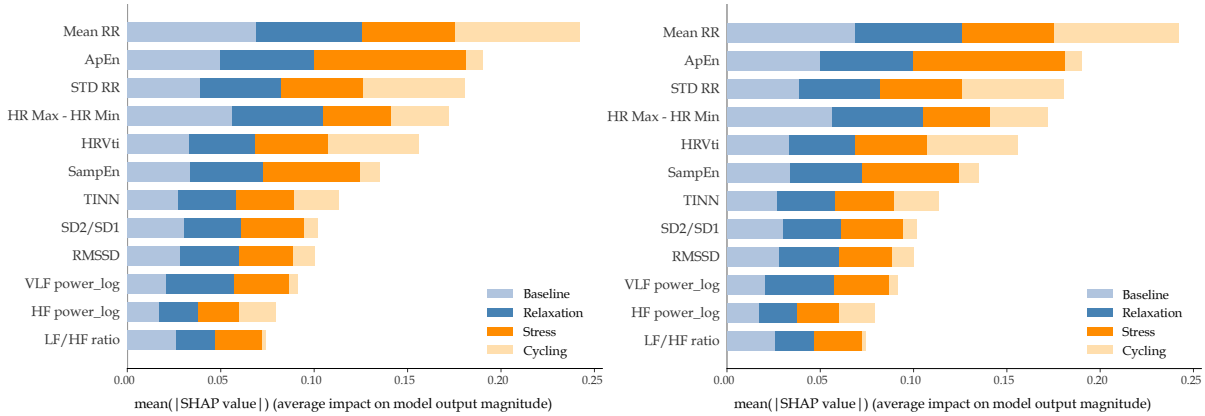
In Figure 10, we present a different analogy, comparing two devices with a single classifier. In this comparison, we have two devices of type ECG and PPG, in Figures 10a, and 10b, respectively. In Figure 10c, we present the effect of multimodality on the model's output. There are differences in the order of feature impact rankings in all three plots. The amount of differences seen in 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 show higher differences as well.

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 11 shows horizontal scatter plots for each feature with different color gradients. Feature importances and feature effects are aggregated in this class-wise summary plot. Each point represents



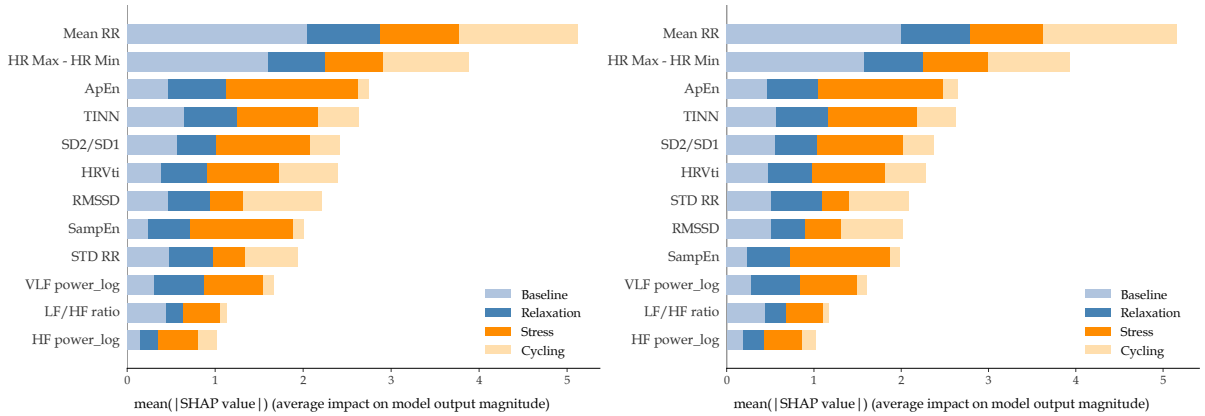
(a) Firstbeat Bodyguard 2 (b) Empatica E4

Figure 11: Feature influences with SHAP for the Stress class, with ExtraTree classifier



(a) Random Forest - Robust scaling

(b) Random Forest - MinMax scaling



(c) LightGBM - Robust scaling

(d) LightGBM - MinMax scaling

Figure 12: Effects of using different types of scalings in model output

a Shapley value for a feature and an observation on this scatter plot. 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 10. ApEn had the 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 Figure 11. Upon close investigation of Figure 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 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. 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 Figures 12c and 12d, 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.

5. Conclusion

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 [20], the end-user's 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 outcome of a model and change the importance of a feature.

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.

One of the limitations in the current study, 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 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.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Niaz Chalabianloo: Methodology, Study Design, Data analysis, Visualization, Writing, and Data acquisition. **Yekta Said Can:** Methodology, and Writing. **Muhammad Umair:** Writing, Study Design, Participant recruitment, and Data acquisition. **Corina Sas:** Study Design, Review and editing, Resources, Project administration, and Funding acquisition. **Cem Ersoy:** Review and editing, Supervision, and Project administration.

Acknowledgement

This work has been supported by AffecTech: Personal Technologies for Affective Health, Innovative Training Network funded by the H2020 People Programme under Marie Skłodowska-Curie grant agreement No 722022.

References

[1] M. R. Kamdar, M. J. Wu, Prism: a data-driven platform for monitoring mental health, in: Biocomputing 2016: Proceedings of the Pacific Symposium, World Scientific, 2016, pp. 333–344.

[2] E. Brondolo, K. Byer, P. Gianaros, C. Liu, A. Prather, K. Thomas, C. Woods-Giscombé, L. Beatty, P. DiSandro, G. Keita, Stress and health disparities: Contexts, mechanisms, and interventions among racial/ethnic minority and low socioeconomic status populations, American Psychological Association (APA) Working Group Report (2017).

[3] D. Colombo, J. Fernández-Álvarez, C. Suso-Ribera, P. Cipresso, H. Valev, T. Leufkens, C. Sas, A. Garcia-Palacios, G. Riva, C. Botella, The need for change: Understanding emotion regulation antecedents and consequences using ecological momentary assessment., Emotion 20 (1) (2020) 30.

[4] O. Van den Bergh, Principles and practice of stress management, Guilford Publications, 2021.

[5] M. Umair, C. Sas, N. Chalabianloo, C. Ersoy, Exploring personalized vibrotactile and thermal patterns for affect regulation, in: Designing Interactive Systems Conference 2021, DIS '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 891–906.

[6] K. M. Richardson, H. R. Rothstein, Effects of occupational stress management intervention programs: a meta-analysis., Journal of occupational health psychology 13 (1) (2008) 69.

[7] C. Bellarosa, P. Y. Chen, The effectiveness and practicality of occupational stress management interventions: A survey of subject matter expert opinions., Journal of occupational health psychology 2 (3) (1997) 247.

[8] K. M. Walsh, B. J. Saab, N. A. Farb, Effects of a mindfulness meditation app on subjective well-being: active randomized controlled trial and experience sampling study, JMIR mental health 6 (1) (2019) e10844.

[9] C. Daudén Roquet, C. Sas, Interoceptive interaction: an embodied metaphor inspired approach to designing for meditation, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–17.

[10] A. Alberdi, A. Aztiria, A. Basarab, Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review, Journal of biomedical informatics 59 (2016) 49–75.

[11] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, B.-H. Koo, Stress and heart rate variability: a meta-analysis and review of the literature, Psychiatry investigation 15 (3) (2018) 235.

[12] M. Alfaras, 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 20 (21) (2020) 5968.

[13] L. Barrios, P. Oldrati, S. Santini, A. Lutterotti, Evaluating the accuracy of heart rate sensors based on photoplethysmography for in-the-wild analysis, in: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare, 2019, pp. 251–261.

[14] A. A. Schuurmans, P. de Looff, K. S. Nijhof, C. Rosada, R. H. Scholte, A. Popma, R. Otten, Validity of the empatica e4 wristband to measure heart rate variability (hrv) parameters: A comparison to electrocardiography (ecg), Journal of medical systems 44 (11) (2020) 1–11.

[15] H. Kinnunen, A. Rantanen, T. Kenttä, H. Koskimäki, Feasible assessment of recovery and cardiovascular health: accuracy of nocturnal hr and hrv assessed via ring ppg in comparison to medical grade ecg, Physiological measurement 41 (4) (2020) 04NT01.

[16] Y. S. Can, B. Arnrich, C. Ersoy, Stress detection in daily life scenarios using smart phones and wearable sensors: A survey, Journal of biomedical informatics 92 (2019) 103139.

[17] Y.-H. Kao, P. C.-P. Chao, C.-L. Wey, Design and validation of a new ppg module to acquire high-quality physiological signals for high-accuracy biomedical sensing, IEEE Journal of Selected Topics in Quantum Electronics 25 (1) (2018) 1–10.

[18] R. Gilgen-Ammann, T. Schweizer, 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 119 (7) (2019) 1525–1532.

[19] M. Murero, 21 wearable internet for wellness and health, Geographies of the Internet (2020) 334.

[20] M. Umair, N. Chalabianloo, C. Sas, C. Ersoy, HRV and Stress: A mixed-methods approach for comparison of wearable heart rate sensors for biofeedback, IEEE Access 9 (2021) 14005–14024.

[21] T. Iqbal, A. Elahi, P. Redon, P. Vazquez, W. Wijns, A. Shahzad, A review of biophysiological and biochemical indicators of stress for connected and preventive healthcare, Diagnostics 11 (3) (2021) 556.

[22] H. G. van Lier, M. E. Pieterse, A. Garde, M. G. Postel, H. A. de Haan, M. M. Vollenbroek-Hutten, J. M. Schraagen, 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 (2019) 1–23.

[23] M. Bolanos, H. Nazeran, E. Haltiwanger, Comparison of heart rate variability signal features derived from electrocardiography and photoplethysmography in healthy individuals, in: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2006, pp. 4289–4294.

[24] N. Selvaraj, A. Jaryal, J. Santhosh, K. K. Deepak, S. Anand, Assessment of heart rate variability derived from finger-tip photoplethysmography as compared to electrocardiography, Journal of medical engineering & technology 32 (6) (2008) 479–484.

[25] V. Jeyhani, S. Mahdiani, M. Peltokangas, A. Vehkaoja, Comparison of hrv parameters derived from photoplethysmography and electrocardiography signals, in: 2015 37th annual international conference of the ieee engineering in medicine and biology society (EMBC), IEEE, 2015, pp. 5952–5955.

[26] P. Konstantinou, A. Trigeorgi, C. Georgiou, A. T. Gloster, G. Panayiotou, M. Karekla, Comparing apples and oranges or different types of citrus fruits? using wearable versus stationary devices to analyze psychophysiological data, Psychophysiology 57 (5) (2020) e13551.

[27] S. Ollander, C. Godin, A. Campagne, S. Charbonnier, A comparison of wearable and stationary sensors for stress detection, in: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2016, pp. 004362–004366.

[28] M. Ragot, N. Martin, S. Em, N. Pallamin, J.-M. Diverrez, Emotion recognition using physiological signals: laboratory vs. wearable sensors, in: International Conference on Applied Human Factors and Ergonomics, Springer, 2017, pp. 15–22.

[29] P. Siirtola, Continuous stress detection using the sensors of commercial smartwatch, in: Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, 2019, pp. 1198–1201.

[30] F. de Arriba Perez, J. M. Santos-Gago, M. Caeiro-Rodríguez, M. J. F. Iglesias, Evaluation of commercial-off-the-shelf wrist wearables to estimate stress on students, JoVE (Journal of Visualized Experiments) (136) (2018) e57590.

[31] A. Greco, G. Valenza, J. Lázaro, J. M. Garzón-Rey, J. Aguiló, C. De-la Camara, R. Bailón, E. P. Scilingo, Acute stress state classification based on electrodermal activity modeling, IEEE Transactions on Affective Computing (2021).

[32] N. Pinheiro, R. Couceiro, J. Henriques, J. Muehlsteff, I. Quintal, L. Goncalves, P. Carvalho, Can ppg be used for hrv analysis?, in: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2016, pp. 2945–2949.

[33] P. K. Stein, Y. Pu, Heart rate variability, sleep and sleep disorders, Sleep medicine reviews 16 (1) (2012) 47–66.

[34] C.-L. Lu, X. Zou, W. C. Orr, J. Chen, Postprandial changes of sympathovagal balance measured by heart rate variability, Digestive diseases and sciences 44 (4) (1999) 857–861.

[35] F. Zimmermann-Viehoff, J. Thayer, J. Koenig, C. Herrmann, C. S. Weber, 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 19 (4) (2016) 169–175.

[36] World Medical Association, World medical association declaration of helsinki: Ethical principles for medical research involving human subjects, JAMA 310 (20) (2013) 2191–2194.

[37] L. Menghini, E. Gianfranchi, N. Cellini, E. Patron, M. Tagliabue, M. Sarlo, Stressing the accuracy: Wrist-worn wearable sensor validation over different conditions, Psychophysiology 56 (11) (2019) e13441.
[38] G. Regalia, F. Onorati, M. Lai, C. Caborni, R. W. Picard, Multimodal wrist-worn devices for seizure detection and advancing research: Focus on the empatica wristbands, Epilepsy Research 153 (2019) 79 – 82.
[39] M. Gjoreski, M. Luštrek, M. Gams, H. Gjoreski, Monitoring stress with a wrist device using context, Journal of biomedical informatics 73 (2017) 159–170.
[40] V. Montesinos, F. Dell'Agnola, A. Arza, A. Aminifar, D. Atienza, Multi-modal acute stress recognition using off-the-shelf wearable devices, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 2196–2201.
[41] Y. Matsumoto, T. Mizuno, K. Mito, N. Itakura, Mental stress evaluation method using photoplethysmographic amplitudes obtained from a smartwatch, in: International Conference on Human-Computer Interaction, Springer, 2021, pp. 357–362.
[42] D. Jaiswal, A. Chowdhury, D. Chatterjee, R. Gavas, Unobtrusive smart-watch based approach for assessing mental workload, in: 2019 IEEE Region 10 Symposium (TENSYMP), IEEE, 2019, pp. 304–309.
[43] A. Shcherbina, C. M. Mattsson, D. Waggott, H. Salisbury, J. W. Christle, T. Hastie, M. T. Wheeler, E. A. Ashley, Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort, Journal of personalized medicine 7 (2) (2017) 3.
[44] E. Weston, P. Le, W. S. Marras, A biomechanical and physiological study of office seat and tablet device interaction, Applied Ergonomics 62 (2017) 83 – 93.
[45] C. Ottaviani, L. Shahabi, M. Tarvainen, I. Cook, M. Abrams, D. Shapiro, Cognitive, behavioral, and autonomic correlates of mind wandering and perseverative cognition in major depression, Frontiers in neuroscience 8 (2015) 433.
[46] B. Crespo-Ruiz, S. Rivas-Galan, C. Fernandez-Vega, C. Crespo-Ruiz, L. Maicas-Perez, Executive stress management: Physiological load of stress and recovery in executives on workdays, International Journal of Environmental Research and Public Health 15 (12) (2018) 2847.
[47] T. Föhr, 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 27 (6) (2017) 612–621.
[48] N. Ahmed, Y. Zhu, Early detection of atrial fibrillation based on ecg signals, Bioengineering 7 (1) (2020) 16.
[49] D. Batista, H. P. da Silva, A. Fred, C. Moreira, M. Reis, H. A. Ferreira, Benchmarking of the bitalino biomedical toolkit against an established gold standard, Healthcare technology letters 6 (2) (2019) 32–36.
[50] A. Hasanbasic, M. Spahic, D. Bosnjic, V. Mesic, O. Jahic, et al., Recognition of stress levels among students with wearable sensors, in: 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), IEEE, 2019, pp. 1–4.
[51] Ö. GÜnaydin, R. B. Arslan, Stress level detection using physiological sensors, in: 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), IEEE, 2020, pp. 509–512.
[52] K. E. Speer, S. Semple, N. Naumovski, 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 10 (1) (2020) 390–404.
[53] M. Umair, N. Chalabianloo, C. Sas, C. Ersoy, A comparison of wearable heart rate sensors for hrv biofeedback in the wild: An ethnographic study, in: 25th annual international CyberPsychology, CyberTherapy & Social Networking Conference, 2020.
[54] C. Chen, C. Li, C.-W. Tsai, X. Deng, Evaluation of mental stress and heart rate variability derived from wrist-based photoplethysmography, in: 2019 IEEE Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS), IEEE, 2019, pp. 65–68.
[55] J. M. Nwaogu, A. P. Chan, Work-related stress, psychophysiological strain, and recovery among on-site construction personnel, Automation in Construction 125 (2021) 103629.
[56] V. Mishra, S. Sen, G. Chen, T. Hao, J. Rogers, C.-H. Chen, D. Kotz, Evaluating the reproducibility of physiological stress detection models, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 4 (4) (2020) 1–29.
[57] D. Miranda, M. Calderón, J. Favela, Anxiety detection using wearable monitoring, in: Proceedings of the 5th Mexican conference on human-computer interaction, 2014, pp. 34–41.
[58] A. V. Bakhchina, K. R. Arutyunova, A. A. Sozinov, A. V. Demidovsky, Y. I. Alexandrov, Sample entropy of the heart rate reflects properties of the system organization of behaviour, Entropy 20 (6) (2018) 449.
[59] T. V. Chernigovskaya, S. B. Parin, I. S. Parina, A. A. Konina, D. K. Urikh, Y. O. Yachmonina, M. A. Chernova, S. A. Polevaya, Simultaneous interpreting and stress: pilot experiment, International journal of psychophysiology 108 (2016) 165.
[60] G. Mark, S. T. Iqbal, M. Czerwinski, P. Johns, A. Sano, Y. Lutchyn, Email duration, batching and self-interruption: Patterns of email use on productivity and stress, in: Proceedings of the 2016 CHI conference on human factors in computing systems, 2016, pp. 1717–1728.
[61] P. A. Afulani, L. Ongeri, J. Kinyua, M. Temmerman, W. B. Mendes, 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 21 (1) (2021) 1–16.
[62] I. T. Hettiarachchi, S. Hanoun, D. Nahavandi, S. Nahavandi, Validation of polar oh1 optical heart rate sensor for moderate and high intensity physical activities, PloS one 14 (5) (2019).
[63] J. Determan, M. A. Akers, T. Albright, B. Browning, C. Martin-Dunlop, P. Archibald, V. Caruolo, The impact of biophilic learning spaces on student success, Architecture Planning Interiors. Architecture Planning Interiors (2019).
[64] H. White, The effects of mindfulness exercises derived from acceptance and commitment therapy during recovery from work-related stress (2019).
[65] S. Laborde, E. Mosley, J. F. Thayer, Heart rate variability and cardiac vagal tone in psychophysiological research - Recommendations for experiment planning, data analysis, and data reporting 8 (FEB) (2017) 213.
[66] F. Scarpina, S. Tagini, The stroop color and word test, Frontiers in Psychology 8 (apr 2017).
[67] J. R. Stroop, Studies of interference in serial verbal reactions., Journal of experimental psychology 18 (6) (1935) 643.
[68] F. de Arriba-Pérez, J. M. Santos-Gago, M. Caeiro-Rodríguez, M. Ramos-Merino, Study of stress detection and proposal of stress-related features using commercial-off-the-shelf wrist wearables, Journal of Ambient Intelligence and Humanized Computing 10 (12) (2019) 4925–4945.

[69] C. Kirschbaum, K.-M. Pirke, D. H. Hellhammer, The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting, Neuropsychobiology 28 (1-2) (1993) 76–81.
[70] P. Brodal, The central nervous system: structure and function, oxford university Press, 2004.
[71] S.-W. Chen, J.-W. Liaw, Y.-J. Chang, L.-L. Chuang, 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 63 (2015) 133–142.
[72] D. Hernando, A. Hernando, J. A. Casajus, P. Laguna, N. Garatachea, R. Bailon, Methodological framework for heart rate variability analysis during exercise: application to running and cycling stress testing, Medical & biological engineering & computing 56 (5) (2018) 781–794.
[73] D. T. Fitch, J. Sharpnack, 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 70 (2020) 81–97.
[74] J.-H. Hong, J. Ramos, A. K. Dey, Understanding physiological responses to stressors during physical activity, in: Proceedings of the 2012 ACM conference on ubiquitous computing, ACM, 2012, pp. 270–279.
[75] J. Parak, I. Korhonen, Accuracy of firstbeat bodyguard 2 beat-to-beat heart rate monitor, White Pap Firstbeat Technol Ltd (2013).
[76] Empatica, medical devices, ai and algorithms for remote patient monitoring.
[77] I. Krumpal, Determinants of social desirability bias in sensitive surveys: a literature review, Quality & quantity 47 (4) (2013) 2025–2047.
[78] R. W. Simon, L. E. Nath, Gender and emotion in the united states: Do men and women differ in self-reports of feelings and expressive behavior?, American journal of sociology 109 (5) (2004) 1137–1176.
[79] M. Malik, Heart rate variability: Standards of measurement, physiological interpretation, and clinical use: Task force of the european society of cardiology and the north american society for pacing and electrophysiology, Annals of Noninvasive Electrocardiology 1 (2) (1996) 151–181.
[80] M. P. Tarvainen, J.-P. Niskanen, J. A. Lipponen, P. O. Ranta-Aho, P. A. Karjalainen, Kubios hrv–heart rate variability analysis software, Computer methods and programs in biomedicine 113 (1) (2014) 210–220.
[81] Y. S. Can, N. Chalabianloo, D. Ekiz, C. Ersoy, Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study, Sensors 19 (8) (2019) 1849.
[82] J. A. Lipponen, M. P. Tarvainen, A robust algorithm for heart rate variability time series artefact correction using novel beat classification, Journal of medical engineering & technology 43 (3) (2019) 173–181.
[83] D. Hernando, S. Roca, J. Sancho, Á. Alesanco, R. Bailón, Validation of the apple watch for heart rate variability measurements during relax and mental stress in healthy subjects, Sensors 18 (8) (2018) 2619.
[84] B. Vescio, M. Salsone, A. Gambardella, 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 18 (3) (2018) 844.
[85] D. J. Plews, B. Scott, M. Altini, M. Wood, A. E. Kilding, 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 12 (10) (2017) 1324–1328.
[86] C. C. Grant, D. C. van Rensburg, N. Strydom, M. Viljoen, Importance of tachogram length and period of recording during noninvasive investigation of the autonomic nervous system, Annals of Noninvasive Electrocardiology 16 (2) (2011) 131–139.
[87] D. Lykken, R. Rose, B. Luther, M. Maley, Correcting psychophysiological measures for individual differences in range., Psychological bulletin 66 (6) (1966) 481.
[88] G. C. Cawley, N. L. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, The Journal of Machine Learning Research 11 (2010) 2079–2107.
[89] S. Varma, R. Simon, Bias in error estimation when using cross-validation for model selection, BMC bioinformatics 7 (1) (2006) 1–8.
[90] D. Krstajic, L. J. Buturovic, D. E. Leahy, S. Thomas, Cross-validation pitfalls when selecting and assessing regression and classification models, Journal of cheminformatics 6 (1) (2014) 1–15.
[91] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.
[92] J. Friedman, T. Hastie, R. Tibshirani, et al., The elements of statistical learning, Vol. 1, Springer series in statistics New York, 2001.
[93] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Machine learning 63 (1) (2006) 3–42.
[94] A. Natekin, A. Knoll, Gradient boosting machines, a tutorial, Frontiers in neurorobotics 7 (2013) 21.
[95] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems 30 (2017) 3146–3154.
[96] S. Raschka, Model evaluation, model selection, and algorithm selection in machine learning, arXiv preprint arXiv:1811.12808 (2018).
[97] M. Hollander, D. A. Wolfe, E. Chicken, Nonparametric statistical methods, Vol. 751, John Wiley & Sons, 2013.
[98] M. Neuhäuser, F. Bretz, Nonparametric all-pairs multiple comparisons, Biometrical Journal: Journal of Mathematical Methods in Biosciences 43 (5) (2001) 571–580.
[99] A. Dinno, Nonparametric pairwise multiple comparisons in independent groups using dunn's test, The Stata Journal 15 (1) (2015) 292–300.
[100] D. G. Kilpatrick, Differential responsiveness of two electrodermal indices to psychological stress and performance of a complex cognitive task, Psychophysiology 9 (2) (1972) 218–226.
[101] P. Sanches, K. Höök, C. Sas, A. Ståhl, Ambiguity as a resource to inform proto-practices: The case of skin conductance, ACM Transactions on Computer-Human Interaction (TOCHI) 26 (4) (2019) 1–32.
[102] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, S. H. A. Chen, Neurokit2: A python toolbox for neurophysiological signal processing, Behavior Research Methods (Feb 2021).
[103] A. Tazarv, S. Labbaf, S. M. Reich, N. Dutt, A. M. Rahmani, M. Levorato, Personalized stress monitoring using wearable sensors in everyday settings, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2021, pp. 7332–7335.
[104] K. M. Dalmeida, G. L. Masala, Hrv features as viable physiological markers for stress detection using wearable devices, Sensors 21 (8) (2021) 2873.
[105] P. Bobade, M. Vani, Stress detection with machine learning and deep learning using multimodal physiological data, in: 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE, 2020, pp. 51–57.

[106] R. K. Nath, H. Thapliyal, A. Caban-Holt, Validating physiological stress detection model using cortisol as stress bio marker, in: 2020 IEEE International Conference on Consumer Electronics (ICCE), IEEE, 2020, pp. 1–5.
[107] R. Lima, D. F. de Noronha Osório, H. Gamboa, Heart rate variability and electrodermal activity in mental stress aloud: Predicting the outcome., in: Biosignals, 2019, pp. 42–51.
[108] A. O. Akmandor, N. K. Jha, Keep the stress away with soda: Stress detection and alleviation system, IEEE Transactions on Multi-Scale Computing Systems 3 (4) (2017) 269–282.
[109] V. Vanitha, P. Krishnan, Real time stress detection system based on eeg signals (2016).
[110] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st international conference on neural information processing systems, 2017, pp. 4768–4777.
[111] L. A. Ferreira, F. G. Guimarães, R. Silva, Applying genetic programming to improve interpretability in machine learning models, in: 2020 IEEE Congress on Evolutionary Computation (CEC), IEEE, 2020, pp. 1–8.
[112] R. C. Deo, Machine learning in medicine, Circulation 132 (20) (2015) 1920–1930.
[113] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
[114] H. Baniecki, W. Kretowicz, P. Piatyszek, J. Wisniewski, P. Biecek, dalex: Responsible machine learning with interactive explainability and fairness in python, Journal of Machine Learning Research 22 (214) (2021) 1–7.
[115] S. Mazzanti, Shap values explained exactly how you wished someone explained to you, Towards Data Science, January 3 (2020) 2020.

Niaz Chalabianloo: received the master's degree in computer engineering from the Middle East Technical University in Turkey, and the PhD degree from Boğaziçi University. Previously, he worked as a researcher in the Marie Sklodowska-Curie H2020 European ITN project AffecTech (Personal Technologies For Affective Health) for three years. He also worked as an adjunct lecturer at the Azad University of Tabriz for two years. His research interests include Affective Computing, Digital Health, Machine Learning, and Ubiquitous Computing.

Yekta Said Can: received the B.Sc, M.Sc. and Ph.D. degrees from Boğaziçi University, in 2012 and 2014 and 2020 respectively. He is currently working on recognizing emotions and stress in Augsburg University as a Postdoctoral Researcher. He also worked as a Teaching Assistant in Bogazici University for six years during his PhD. His research interests include biometrics, physiological signal processing, affective and wearable computing, social signal processing and machine learning.

Muhammad Umair: is an Innovation Fellow at Open Lab, Newcastle University. He is interested in co-creative approaches to design research, and exploring the potential of tangible interfaces and data-driven systems to support health and well-being in an accessible way. Previously, he worked on the design and evaluation of tangible technologies to support awareness and regulation of affect. This included using biofeedback i.e., sensing data from sensors on the body, and providing feedback using non-screen-based technologies (vibration, shape-changing, temperature) to support awareness and regulation of affect.

Corina Sas: is a Professor of Human-Computer Interaction and Digital Health with the School of Computing and Communications, and Assistant Dean for Research Enhancement at Lancaster University, UK. Her research is in the area of technologies for well-being and health. She published over 200 papers, and her work received extensive media coverage as well as five Best Paper and Honourable Mention Awards. She has been an investigator on grants totaling over £15.1 million, and coordinated the AffecTech: Personal Technologies for Affective Health, a Marie Sklodowska-Curie Innovative Training Network (2017-2021) funded by the European Commission. Prof Sas is also part of the Editorial Boards of the ACM Transactions in Human-Computer Interaction, and Taylor Francis Human Computer Interaction journals. She received a Ph.D. degree in Human-Computer Interaction from University College Dublin.

Cem Ersoy: received the Ph.D. degree from Polytechnic University, New York, NY, USA, in 1992. He was a Research and Development Engineer with NETAŞ, from 1984 to 1986. He is currently a Professor and the computer engineering department head with Boğaziçi University. He is also the Vice Director of the Telecommunications and Informatics Technologies Research Center, TETAM. His research interests include wireless/cellular/ad-hoc/sensor networks, activity recognition, and ambient intelligence for pervasive health applications, green 5G and beyond networks, mobile cloud/edge/fog computing, software-defined networking, and infrastructure-less communications for disaster management. He is a member of IFIP. He was the Chairman of the IEEE Communications Society Turkish Chapter for eight years.