IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 16, 2020, accepted January 5, 2021, date of publication January 18, 2021, date of current version January 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3052131

HRV and Stress: A Mixed-Methods Approach for Comparison of Wearable Heart Rate Sensors for Biofeedback

MUHAMMAD UMAIR[1], NIAZ CHALABIANLOO[2], CORINA SAS[1], AND CEM ERSOY[2], (Senior Member, IEEE)

[1]School of Computing and Communications, Lancaster University, Lancaster LA1 4YW, U.K.
[2]Department of Computer Engineering, Boğaziçi University, Istanbul 34342, Turkey

Corresponding author: Niaz Chalabianloo (niaz.chalabianloo@boun.edu.tr)

*Muhammad Umair and Niaz Chalabianloo contributed equally to this work.

This work was supported by the AffecTech: Personal Technologies for Affective Health, Innovative Training Network funded by the H2020 People Programme under Marie Skłodowska-Curie under Grant 722022.

ABSTRACT Stress is one of the most significant health problems in today's world. Existing work has used heart rate variability (HRV) to detect stress and provide biofeedback in order to regulate it. There has been a growing interest in using wearable biosensors to measure HRV. Each of these sensors acquires heart rate data using different technologies for various bodily locations, therefore posing a challenge for researchers to decide upon a particular device in a research experiment. Previous work has only compared different sensing devices against a gold standard in terms of data quality, thus overlooking qualitative analysis for the usability and acceptability of such devices. This paper introduces a mixed-methods approach to compare the data quality and user acceptance of the six most common wearable heart rate monitoring biosensors. We conducted a 70-minute data collection procedure to obtain HRV data from 32 participants followed by a 10-minute semi-structured interview on sensors' wearability and comfort, long-term use, aesthetics, and social acceptance. We performed quantitative analysis consisting of correlation and agreement analysis on the HRV data and thematic analysis on qualitative data obtained from interviews. Our results show that the electrocardiography (ECG) chest strap achieved the highest correlation and agreement levels in all sessions and had the lowest amount of artifacts, followed by the photoplethysmography (PPG) wristband, ECG sensor board kit and PPG smartwatch. In all three sessions, wrist-worn devices showed a lower amount of agreement and correlation with the reference device. Qualitative findings from interviews highlight that participants prefer wrist and arm-worn devices in terms of aesthetics, wearability, and comfort, followed by chest-worn devices. Moreover, participants mentioned that the latter are more likely to invite social judgment from others, and they would not want to wear it in public. Participants preferred the chest strap for short-term use and the wrist and arm-worn sensors over long-time.

INDEX TERMS Affective computing, biofeedback, biosensors, HRV, physiological signals, stress, wearable sensors.

I. INTRODUCTION

Stress is one of modern life's primary mental health concerns. It is our body's reaction to unmanageable pressure from a situation or life event resulting in emotional, behavioral, or bodily changes [1]. Long-term exposure to stress can cause adverse outcomes both on mental and physical health, such as depression, anxiety, and cardiovascular diseases [1]–[3]. In Great Britain alone, work-related stress, depression, or anxiety accounted for 44% of work-related ill health cases in 2018/19 [4]. The ever-growing problems caused by psychological stress and its severe negative consequences on public health can subsequently inflict social, economic, and financial burden on societies [3], [5]. To overcome these problems, researchers in various disciplines have developed biofeedback systems to help end-users manage stress [6]–[10]. Heart rate variability (HRV) biofeedback in particular has been commonly used to visualize users' internal states that are hard to experience otherwise in order to guide the user to control physiological activities related to stress and anxiety

The associate editor coordinating the review of this manuscript and approving it for publication was Venkata Rajesh Pamula.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

disorders [11]–[15]. HRV provides regulation of autonomic balance and can be measured from a single sensor [11], [16].

Increasing work on HRV assessment has used a variety of mobile, mostly wearable inconspicuous sensing devices. These devices use different technologies [17], and are designed to be placed on different bodily locations [17]–[19]. Moreover, such devices produce HRV data varying in quality and also differ in user acceptance based on wearability, aesthetics and ease of use. The quality of data sensing and its user acceptance are important factors when deciding to choose a sensor for HRV monitoring. Existing work compared different sensing methodologies and argued for either one or a combination of both [20], [21], however the majority of such work compared a particular sensing device to a golden standard reference device only performing quantitative analysis for assessing the correlation and agreement between methods of measurement [22]–[24], while only a few perform qualitative analysis for usability and acceptability [25].

Considering the equal importance of data quality and user acceptance of the HRV sensing devices, this paper extends prior work by introducing a mixed-methods approach consisting of both data quality and user acceptance. Combining qualitative and quantitative methods can yield greater depth and breadth of information, which is often not possible while applying a singular approach [26], [27]. We argue that combining quantitative and qualitative data would help unpack holistic understandings of the critical factors involved in choosing a particular HRV monitoring device. This paper aims to explore HRV data quality in respect to its features and user acceptance of key HRV sensors. Specifically, we address the following research questions:

- How do HRV measurement sensors differ in features (RMSSD, pNN50, Mean RR, HF, LF and LF/HF) in terms of correlation and agreement levels compared to a reference device?
- What are users' views and experiences in terms of wearability and comfort, long-term use, aesthetics, and social acceptance of these sensing devices?

In order to address these research questions, we combined quantitative and qualitative data analysis techniques across six heart rate monitoring devices on key bodily locations. All of which are widely used in the literature for HRV analysis [17], [28]–[32] and have been validated in prior research [33]–[35]. In total, we recruited 32 participants and asked them to wear six different heart rate monitoring devices. All participants underwent an individual data collection session, including Baseline, Stress, and Relaxation stages. We performed quantitative analysis consisting of the most common and well-known agreement and correlation tests on the HRV data collected from five of these heart rate monitors. In particular, to study the destructive influences of artifacts on HRV data quality and the influence of proper artifact removal thresholds on the effective recovery of the noisy data, we applied multiple levels of artifact removal thresholds, i.e. (automatic, medium, and strong) on all of the data from five HRV monitoring devices followed by Pearson and Spearman's correlation and Bland-Altman agreement analysis. Moreover, we conducted semi-structured interviews with participants and performed thematic analysis on the interview data for extracting participants' views and experiences on the sensors' wearability and comfort, long-term use, aesthetics, and social acceptance.

The rest of the paper is structured as follows. Section II provides a review of related literature on HRV and biofeedback and outlines the studies conducted for the comparison of HRV data. Section III describes our data collection method and sensing devices used to collect HRV data. Section IV presents the data analysis steps and details both qualitative and quantitative results. Finally, Section V closes with conclusions and future directions.

II. RELATED WORK

Our work is inspired by two strands of existing literature, namely: HRV and biofeedback, and comparison and validation studies for HRV.

A. HRV AND BIOFEEDBACK

Affective health and wellbeing is a growing area of research. Many have developed affective technologies targeting stress-related disorders for adults, children, caregivers, clinical staff, and researchers [36], [37]. Such technologies aim to provide prevention, diagnosis, triage, intervention, self-management, and maintenance of affective disorders in both clinical and non-clinical settings. An essential function of interactive systems for affective health is HRV biofeedback.

HRV is a measure of both sympathetic and parasympathetic branches of the autonomic nervous system (ANS). When the heart beats, it triggers an electrical impulse and blood flow in the body, which can be captured by biosensors. These sensors measure heart rate in beats per minute (BPM), which are not always at a constant frequency. HRV is a non-invasive and straightforward measurement representing the changes in the time interval between successive heartbeats [16]. These variations in heartbeats are measured in milliseconds and are called interbeat interval (IBI), RR intervals, or NN intervals. HRV can be analyzed using time and frequency domain methods [38], [39]. Common time-domain features include mean RR (mean value of the inter-beat intervals), STD RR (standard deviation of the inter-beat intervals), RMSSD (root mean square of successive differences of the RR intervals) and pNN50 (percentage of successive beat-to-beat intervals that differ by more than 50 ms). Frequency-domain features include LF (power in the low-frequency band), HF (power in the high-frequency band), and LF/HF (ratio of low-frequency to high-frequency).

HRV is influenced during activities such as exercise, eat, and sleep. Moreover, it is closely related to emotional arousal and decreases when an individual is emotionally stressed. This is of special significance considering the involvement of the parasympathetic activity or otherwise called vagal tone, with processes of self-regulation connected to

TABLE 1. Heart rate variability features and their definitions.

Feature	Description
Mean RR	Mean value of the inter-beat (RR) intervals
STD RR	Standard deviation of the inter-beat interval
RMSSD	Root mean square of successive differences of the RR intervals
pNN50	Percentage of the number of successive RR intervals varying more than 50ms from the previous interval
SDSD	Related standard deviation of successive RR interval differences
TINN	Triangular interpolation of RR interval histogram
LF	Power in low-frequency band (0.04-0.15 Hz
HF	Power in high-frequency band (0.15-0.4 Hz)
LF/HF	Ratio of LF-to-HF
pLF	Prevalent low-frequency oscillation of heart rate
pHF	Prevalent high-frequency oscillation of heart rate
VLF	Power in very low-frequency band (0.00-0.04 Hz)

emotional, affective, and psychological well-being [40]. Different circumstances can influence both LF and HF. Existing work on HRV components suggests that HF is affected by parasympathetic activity [41], [42], whereas LF is believed to be reflecting both sympathetic and vagal activity [38]. Boonnithi *et al.* investigated time-domain and frequency-domain features of HRV and found mean RR, LF, and LF and HF differences to be the most distinctive features for detecting stress [43]. Time-domain metrics such as pNN50 and RMSSD are more closely correlated with parasympathetic activity than the standard deviation of NN intervals (SDNN) [39]. Some of the most popular HRV features used in the stress detection and biofeedback studies are listed in Table 1. In addition to measuring regulatory efficiency, HRV can also be used to help rebuild regulatory flexibility through biofeedback [16].

Biofeedback is a technique that uses biosensors to capture internal bodily processes and provide feedback for individuals to actively control complex physiological activities [44]. Examples include systems for self-awareness and reflection [45], [46], emotion regulation [47], [48], as well as those emphasizing the role of the body for relaxation or mindfulness practices [49], [50]. In particular, HRV biofeedback has been used in a variety of disorders and applications because of its effectiveness [11], [51]. Existing work on stress biofeedback pointed to a decrease in SDNN during stress and increase in RMSSD during biofeedback-assisted breathing [10].

HRV biofeedback can be provided through a range of modalities. Most commonly, the feedback is given to the user through visual [48], [52], auditory [53], or haptic changes [45], [47], and has been studied to regulate bodily functions and improve well-being in various stress-related and affective conditions [12], [54]. As presented in Figure 1, HRV biofeedback can be delivered visually either through a smartwatch or a mobile phone screen, whereas the haptic feedback can be provided through a vibrotactile actuator present in some devices. If the feedback modality is not present in

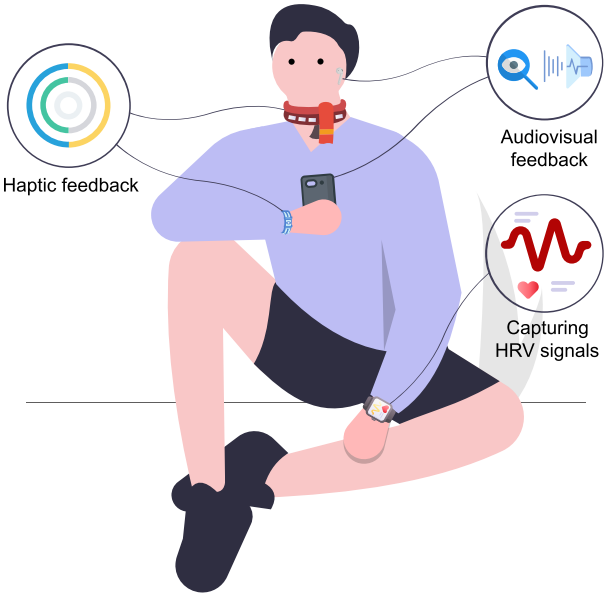


FIGURE 1. Capturing HRV signals and presenting the feedback through visual, auditory or haptic modalities.

the device, an external actuator can be added to deliver the HRV biofeedback [45]. In order to deliver accurate HRV biofeedback, the sensor data must be reliable, and the device should be acceptable for the user in terms of unobtrusiveness, wearability and comfort.

B. COMPARISON AND VALIDATION FOR HRV MONITORING DEVICES

This section outlines the quantitative and qualitative studies conducted for the comparison of HRV data quality and usability of the ubiquitous wearable devices used for the detection of the HRV signals.

1) QUANTITATIVE ANALYSIS

Initial experimental practices in the laboratory settings use medical-grade equipment as the reference golden standard. These devices provide ground truth data and serve as the initial building blocks of comparative studies for evaluating the usability, integrity, and trustworthiness of the ever-growing wearable biosignal monitoring systems.

Investigation of photoplethysmography (PPG) accuracy as an estimate of HRV and the possibility of employing it as a surrogate of electrocardiography (ECG) have been the topics of many experimental and survey studies [19], [21], [55], [56]. In a pioneering study to evaluate the feasibility of utilizing PPG devices for HRV monitoring and the effects of movement on the quality and accuracy of PPG readings, Gil *et al.* concluded that PPG devices presented no statistically significant differences, and there was a high positive correlation with the ECG reference device [56]. Lin *et al.* examined the PPG and ECG recordings from eight healthy subjects before and after exercise. Their findings showed that the HRV obtained from PPG agrees nearly well with the HRV

determined by the ECG signals [57]. They also showed that for the analysis of frequency-domain features of the HRV, it is essential to have at least three minutes of cardiac biosignal recordings and the HRV power spectrum distributions of the data duration of three minutes were similar to the ones with five minutes duration. Finally, they reported that the correlation between PPG and ECG-based HRV power spectrum distribution was observed to be proper for subjects at rest and is decreased following physical exercises.

Binsch *et al.* tested three different wristbands equipped with PPG sensors for the study of heart rate and step count measurement and compared their readings with the ground truth. They concluded that wearable PPG wristbands used in their research could be considered as reliable tools for measuring the heart rate while the subjects under study are at resting conditions. On the other hand, when there are body movements on tasks that require more physical activity, sensor readings become more inaccurate and less reliable [58]. Ge *et al.* compared the accuracy of heart rate measurements of two off the shelf devices, i.e. an Apple watch equipped with a PPG sensor and a Polar Chest strap, which measures the heart rate using ECG signals collected from 50 healthy subjects [59]. Based on their experiment results, in normal condition, when the subjects were at resting state and not asked to perform any physical activity, results were almost similar with a maximum error of around 2%. However, there were variations around 10% between the two device readings during vigorous exercises like walking. In general, the chest brace using ECG tends to be more precise than the Apple Watch, which uses PPG for cardiac monitoring during exercise.

Ollander *et al.* performed a comparison study on both time and frequency domain HRV features calculated from the Empatica E4 and a reference ECG device during a stress detection elicited using the Trier Social Stress Test (TSST). The IBI acquired from Empatica E4 showed a considerable amount of loss, particularly when the subject was asked to perform a task. Nevertheless, the HRV time-domain features, which are always among the top features used in stress detection studies, were measured accurately enough for stress detection [60].

In a recent study, Mejía-Mejía *et al.* investigated the difference between accuracy and quality of HRV measurement using ECG and PPG devices. They conducted a whole-body cold exposure while recording PPG readings from different body locations. They employed Bland-Altman, and analysis of variance to conclude that PPG not only reacts differently to the cold exposure compared to ECG, but also reacts differently among different body parts [19].

2) QUALITATIVE ANALYSIS

Besides the quantitative methods of analysis, existing research has commonly employed qualitative analysis techniques [61]. The qualitative analysis consists of interviews involving study participants' views, opinions, or experiences transcribed in words. Interviews can be of a structured or semi-structured type. In structured interviews, the interviewer follows a set of pre-determined questions, whereas, in semi-structured interviews, the interviewer is free to probe beyond specified questions. Qualitative data are commonly analyzed using thematic analysis [62], which consists of two types of approaches. A typical analysis approach consists of coding the qualitative data and generating themes. Coding means highlighting text from the document, labeling them to describe their content, and later merging them to generate different themes. Thematic analysis can take an inductive or deductive approach. An inductive approach relies on data to drive the themes, whereas a deductive approach means finding out preconceived themes based on existing theories or knowledge [63]. Qualitative data analysis can be performed using any of the different software tools [64], [65], which support the generation of codes and development of themes. Key examples of research on affective biofeedback interfaces using qualitative analysis include Affective Diary [66] to identify bodily experiences, Affective Health [67], [68] to manage stress, and Affective Chronometry to reflect on and regulate affective experiences [45].

The decision to choose an HRV monitoring device depends on a variety of criteria. Some of the factors influencing the efficiency and quality of the device are desirable in all applications, e.g., long-term use is always one of the coveted requirements regardless of the application type. On the other hand, there are other factors that solely depend on the type of target application. While achieving results similar to the gold standard in terms of accuracy is essential in most research-oriented applications, achieving such accuracy levels is not always essential. Certain features like lower weight and size, ease of long-term use and even fancy designs are among the most desired and demanded functionalities in daily life usages [25]. Simonnet and Gourvennec [69] had 11 participants to wear different types of heart rate sensors, smartwatches, chest belts, and ECG electrodes for 24 hours and fill questionnaires regarding the acceptability of the devices. Using this data, devices were compared in terms of acceptability, usability and how data reliability can affect their acceptability.

Thus, we believe that a comparison study between HRV monitoring devices must include both equally important quantitative and qualitative features. Our paper contributes to existing work by providing an in-depth comparison of the six most common heart rate monitoring devices for multiple locations on the body. In addition to quantitative analysis of HRV features, our work also highlights users' experiences and opinions on wearability and comfort, long-term use, device aesthetics, and social acceptance. The primary purpose of our study is to review these devices, allowing the researchers to decide on choosing the most suitable HRV monitoring device for the relevant research.

III. METHODOLOGY

Our study method consisted of a 70-minute session composed of three main stages, as shown in Figure 2. Stage one of

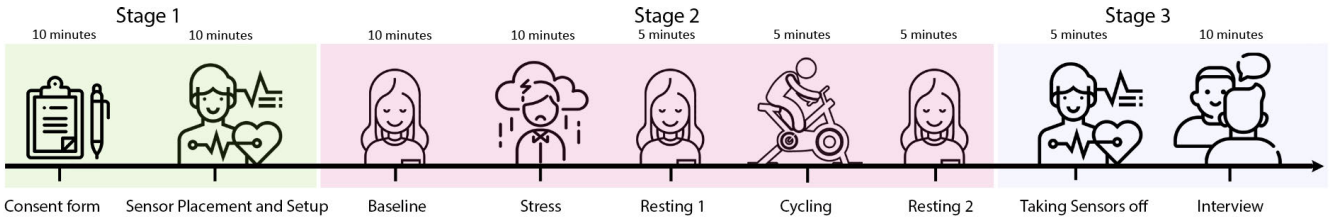


FIGURE 2. **A single session individual data collection procedure followed by participants.** Note: Data acquired in *Cycling* and *Resting 2* phases are not analysed in the study.

TABLE 2. **Heart monitoring sensors along with their placements used during data collection activity.**

Article	Device	Sensors	Sampling rate	Placement	Connectivity	Realtime streaming	Cloud storage	Actuator/Display
[22], [32], [70]	Empatica E4	PPG, EDA, ACC, IR Thermopile	64 Hz	Wrist	Bluetooth	✓	✓	✗
[17], [71]	Samsung Gear S2	PPG, ACC, Barometer, Gyro	100 Hz	Wrist	Bluetooth	✓	✗	Display
[72], [73]	Firstbeat Bodyguard 2	ECG, ACC	1000 Hz	Chest	USB	✗	✗	✗
[74], [75]	BITalino (r)evolution	ECG, EEG, EDA, EMG, ACC	1000 Hz	Chest	Bluetooth	✓	✗	Vibration, Buzzer, Led
[76], [77]	Polar H10	PPG	1000 Hz	Chest	Bluetooth	✓	✓	✗
[78]	Polar OH1	PPG	1 Hz	Arm	Bluetooth	✓	✓	✗

the study included participants' consent, followed by sensors placement and setup. In the second stage, we collected HRV data from Baseline, Stress, Resting 1, Cycling, and Resting 2 using sensors on different bodily locations. All the sensors were taken off in the final stage, followed by a semi-structured interview based on participants' experiences with sensors.

A. SENSORS

This section introduces the sensors employed for the heart rate data acquisition and the settings and features in the software tools used to preprocess and analyze the data. We used six different commercially available heart rate monitoring devices, i.e., Empatica E4, Samsung Gear S2, Firstbeat Bodyguard 2, BITalino (r)evolution board, Polar H10, and Polar OH1 as shown in Table 2.

The first ECG device we selected and chose as the reference device in our experiment is Firstbeat Bodyguard 2. This is a lightweight, reliable, and validated [33] heart rate RR monitor, and yet at the same time, a robust tool for recording the R-R interval in short and long-term durations. Firstbeat Bodyguard 2 automatically starts capturing the data once connected to the skin using two disposable precordial electrodes. ECG signals sampled at 1000 Hz are processed inside the device, and the RR intervals are recorded in the form of an offline output data accessible via the Firstbeat software through a USB connection. The second heart rate monitor used in our study is the Polar H10 chest strap, which is also equipped with ECG sensors and measures heart rate by a 1000 Hz sampling frequency providing the actual RR values instead of raw ECG signal. The Polar H10 can be paired with Android and iOS devices for real-time RR recording with off the shelf and mostly open-source HRV recording applications. However, in order to achieve the most accurate results and eliminate the possibility of data loss or degraded accuracy by using third-party software, we utilized the Polar H10 to transmit its RR data to the Polar V800 sports watch and heart rate activity monitor. In order to monitor the heart rate and record the RR values, the Polar V800 requires a standalone chest strap. In [79], Giles *et al.* conclude that the Polar V800 watch shows RR interval recordings consistent with a medical-grade ECG holter. Upon connecting the Polar V800 to the computer, RR recording results are extracted automatically and synced to the cloud-based Polar flow web service, offering multiple options such as downloading the RR sessions and viewing the summary statistics. The third ECG device in our experiment is a kit from PLUX Wireless Biosignals. BITalino (r)evolution board kit is equipped with multiple types of sensors and actuators, and it comes with a free software called OpenSignals (r)evolution. This tool is required for initialization, connection, and setting up the device. Raw ECG signal acquisition is also performed live via Bluetooth connection using the same software.

The PPG Devices in our study are Empatica E4, Samsung Gear S2, and Polar OH1. Empatica E4 has a battery life of 32+ hours and charges in less than two hours. It can also hold up to 60 hours of data, and besides a USB connection, it is capable of real-time data transfer using Bluetooth. These numbers are a significant improvement over traditional smartwatches in continuous PPG recording mode since Empatica E4 is exclusively designed for research purposes. Empatica E4 records the raw BVP (Blood Volume Pulse) signal from its PPG sensor with a sampling rate of 64 Hz [80]. Diastolic points in the blood volume pulse, which corresponds to the local minima of the BVP signals, are used to compute the inter-beat interval values. Samsung Gear S2 is a smartwatch capable of producing IBI values with PPG sensing. In order to extract the IBI data from the Samsung Gear S2, we developed a smartwatch application for Samsung's Tizen OS. Our application allows the selection of sensors to be employed for data acquisition. When the PPG sensor is selected for continuous recording, Gear 2's battery

provides a maximum run-time of around three hours. Participants were also asked to wear heart monitoring armband, Polar OH1. The Polar OH1 is an arm worn device equipped with a PPG sensor for heart rate monitoring. Its accuracy for heart rate monitoring is validated in [78], and the results showed good agreement with the reference device. Unfortunately, Polar does not provide the RR interval data extraction option in its OH1 armband. Furthermore, the extraction of raw PPG data is also not feasible, making it impossible to utilize the device for HRV analysis; therefore, we do not perform its quantitative analysis and only include it in qualitative analysis involving usability and user acceptance.

In order to guarantee quality data acquisition and avoid the artifact mostly caused by movements and improper sensor to skin contact, medical-grade Ag/AgCl (silver/silver-chloride) self-adhesive electrodes were used by preparing a clean and dry surface for ECG devices. Furthermore, researchers advised all participants to wear the devices in accordance with guidelines given to them and avoid any sudden and unnecessary movements during the data acquisition. For instance, researchers checked whether proper skin contact was established in PPG devices, and devices were not too loose or tight and uncomfortable for participants.

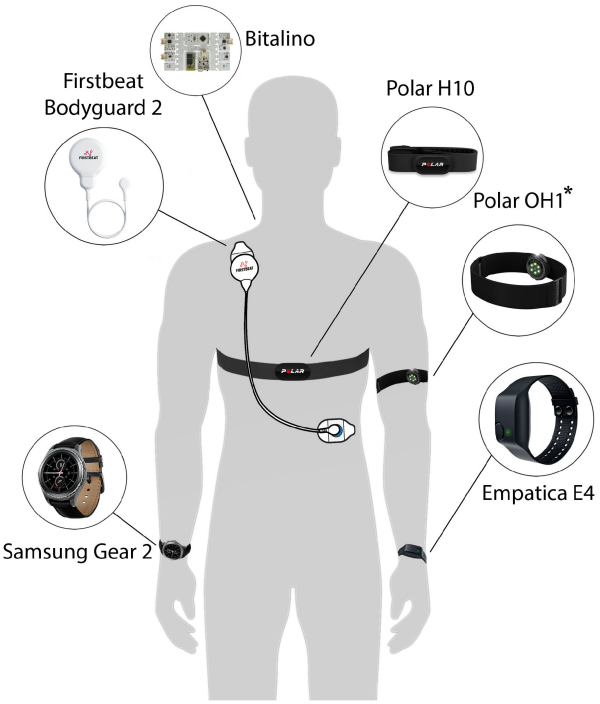


FIGURE 3. All of the heart monitoring sensors connected to subjects for the simultaneous data acquisition. *: Polar OH1 was only included in qualitative analysis.

B. PROCEDURE

The ethics review board of Boğaziçi University had approved the study before we started the data collection sessions. While conducting the data acquisition sessions, all participants underwent identical individual sessions during a single visit to our lab. Before starting the data collection procedure, participants signed an informed consent form. Then, they were given an information sheet that introduced them to the study procedures in detail. Participants were told that their participation is voluntary, and they are free to withdraw at any time during the study and from a week after they have taken part, and all the data related to them used in the study will be destroyed. Participants were excluded from the study if they reported any significant mental and physical conditions, i.e., depression, anxiety, or cardiovascular disease. Each participant was then assigned a unique identification number, and all the relations between the participant names and numbers were anonymized.

At the start of the study, participants were instructed on how to wear all the sensors detailed in the section above. The first two authors conducted the study, assisting participants with the placement of the sensors on different bodily locations for data recording, as shown in Figure 3. The study was conducted in a soundproof room with no visual or auditory distractions. Participants were asked to sit on a chair facing a small clean table towards the wall. The room was set up with two laptops on another table behind participants for data collection to be used by researchers. This whole procedure took about 15-20 minutes and helped participants in getting used to the environment, as shown in Figure 2 (stage 1). In order to achieve clear and precise interpretations of vagal tone and psychophysiological phenomena, it is recommended to avoid movements during measurements [40]. In the majority of studies, baseline recordings are recorded while the subject is seated, and it is recommended that the body position in stressor and relaxation sessions must be as close as possible to the baseline [40]. Accordingly, for data acquisition sessions in our study, we followed the same standards and participants were asked to sit in a comfortable posture with their hands on the table and avoid any significant movements after the sensor setup was ready. The study setup stayed the same for each participant, and facilitators did not take turns during data collection sessions.

C. TASKS AND LOGGED DATA

The study procedure was inspired by similar tasks described in previous work consisting of a Baseline stage and a sequence of short Stressor events followed by a Resting period also called post-event [40]. To collect the baseline data, participants completed an initial resting period while sitting still for ten minutes. Time synced heart rate data from each device was collected for each participant throughout this initial data acquisition procedure.

The next task included participants having to complete stressor tasks for ten minutes, followed by five minutes of each of the following activities: Resting 1, Cycling, and Resting 2, as depicted in Figure 2. The stressor task was composed of two phases. In phase one, participants were introduced to the Stroop Color Test to be performed on a tablet. Stroop is a color and word neuropsychological test extensively used for both experimental and clinical purposes [81]. The Stroop Color Test is based on the fact that

humans are able to read words considerably quicker than they can recognize and identify colors. In the Stroop Color Test, color names (e.g., red, yellow or green) are printed with a different color (e.g., the word ''green'' in red, rather than in green). Once a participant is asked to name the color of a word, if the displayed/printed color of the word does not match the color name, the processing of that particular feature (word) hinders the concurrent processing of the second feature (color), and it takes more time and effort making the participants more vulnerable to make mistakes. This test has been widely used previously as a mental stressor and to study the psycho-physiological responses to mental stress [82]. Participants were asked to perform the test for five minutes and were told that their scores would be compared against other participants. The Stroop test was followed by phase two of the stressor task consisting of arithmetic questions, which is a component of the well-known stress protocol Trier Social Stress Test (TSST) [83]. Participants were asked to count backward with varying differences for five minutes (e.g., counting backward in steps of 17 from any given 4-digit number), while one of the researchers was pretending to take notes of right and wrong answers. For physical stressor, participants were asked to perform cycling, starting with low (60W resistance), medium (90W), and intense cycling exercise (120W) for five minutes. After both mental and physical stressors, a five minute sitting rest period was implemented during which data were continuously acquired. The data acquisition was stopped after the final resting period following physical exercise, and one of the researchers helped participants to take off all the sensors.

The final part of the study included 10 minutes of a semi-structured interview where participants were asked questions on wearability and comfort, long-term use, aesthetics, and social acceptance of the sensors and followed by a 5-point Likert scale questionnaire on these parameters.

Each participant was rewarded with the equivalent of $20 Amazon gift card. Although our data collection procedure includes both mental, i.e., Stroop and TSST and physical stressors data, i.e., cycling, in this paper, we only analyze and provide results on the data collected during Baseline, Stress and Resting 1 phases leaving the data collected from Cycling and Resting 2 phases for the future work.

D. PARTICIPANTS

To recruit participants, we advertised the study through mailing lists and flyers on the campus. Interested participants contacted us through emails and agreed upon a suitable time for participation. Participants were told to follow a normal sleep routine the night before and abstain from caffeinated drinks and meal two hours before the study. In total, we recruited 32 healthy participants (22 Males and 10 Females, age = 28.4±5.98 years, BMI = 25.61±6.49).

IV. DATA ANALYSIS AND RESULTS

This section provides quantitative analysis consisting of data-prepossessing, correlation and agreement analysis on the HRV data followed by qualitative analysis on interview data related to participants' views and experiences on the sensors' wearability, comfort, aesthetics, social acceptance and long-term use as shown in Figure 4.

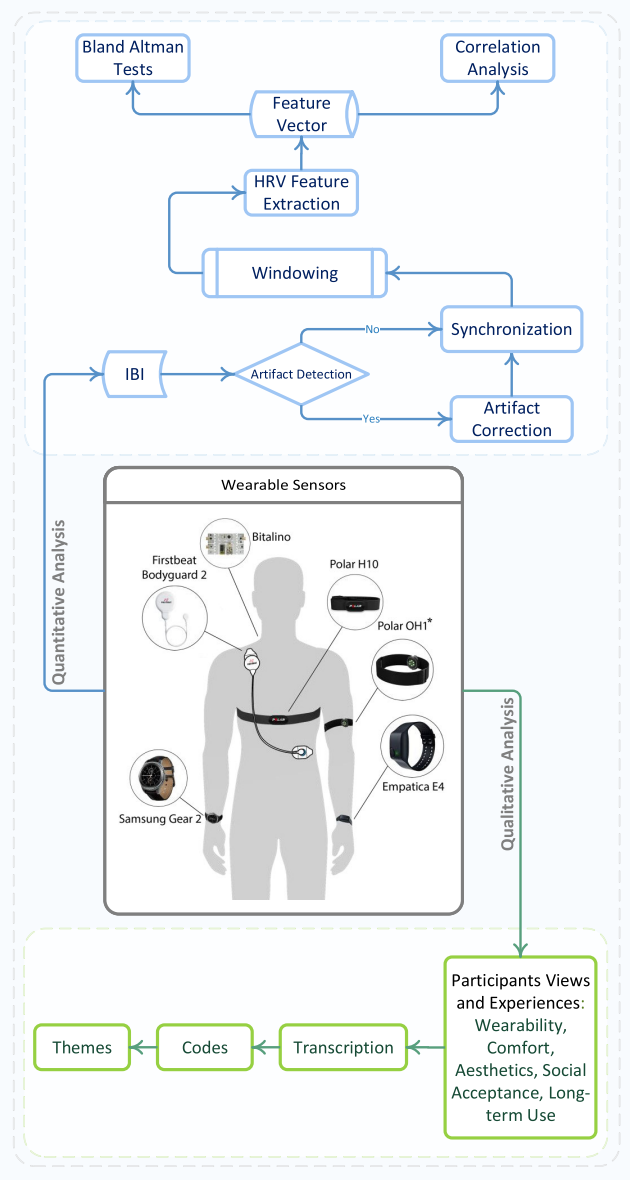


FIGURE 4. Mixed-methods approach for comparison of wearable heart rate sensors. *: Polar OH1 was only included in qualitative analysis.

A. DATA PREPROCESSING

The task force of the European Society of Cardiology and the North American Society for Pacing and Electrophysiology have recommended five minutes as a standard duration of short-term HRV recordings [38]. Moreover, others have investigated even lower than five minutes long RR intervals, also called ultra-short-term recordings (3 minutes, 2 minutes, 1 minute, 30 seconds) [84] and compared their reliability to short-term HRV. They found that the ultra-short-term analysis

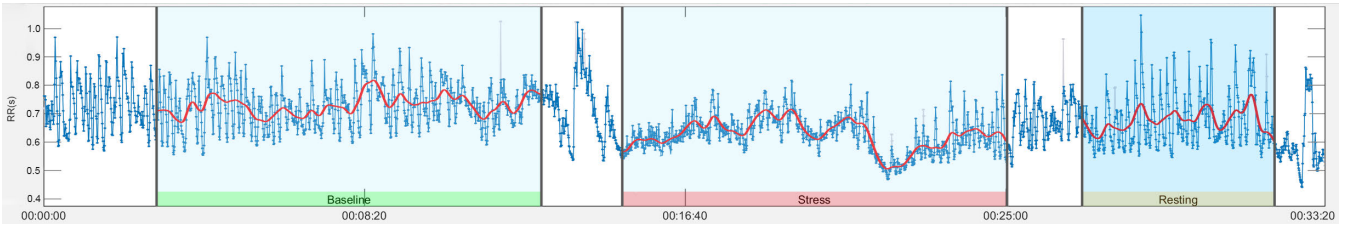


FIGURE 5. Sample RR intervals with ten and five minutes windows recorded from a participant. Note that the gaps between windows are the time between activities. Context is labeled accurately during the data acquisition.

is likely to be a good substitute for the standard five-minute analysis of HRV [84]–[86]. In this paper, we perform HRV analysis on basic (10-minutes), and short-term (5-minutes) recordings.

Before preprocessing the data, we separated the raw heart rate data into three consecutive sessions, i.e., Baseline, Stress, and Resting based on timestamps collected during the data acquisition session, as shown in Figure 5 followed by signal synchronization. In order to preprocess the data acquired from all biosensors, we used a scientifically validated HRV analysis tool, "Kubios" [87] version 3.3. Kubios HRV is currently the most widely used robust HRV analysis tool in the research community, which can be used to analyze the cardiovascular health data or to measure the impact of stress on human health and well-being.

Signal synchronization is a critical step in comparing signals of the same nature recorded from multiple sources to ensure no time drifts between the recordings from the reference and the comparison devices. We performed both automatic and manual data synchronization during data preprocessing to ensure the most precise data alignment. One of the widely used methods in signal synchronization is cross-correlation [18], [23], [88]. We used cross-correlation to find the time shift between the signals from the reference and comparison devices. In this method, the maximum of cross-correlation matches the time-point in which signals are best synchronized. Besides, we compared, reviewed, and synchronized the data by visual inspection of the signals in the Kubios environment and aligning the raw peaks manually in Kubios's data browser. This procedure is conducted in previous studies [34], [89], and we found it to be as accurate as of the cross-correlation method.

1) RR DETECTION

Detection of the QRS values from the ECG signals and pulse waves from the PPGs are accurately performed by Kubios HRV software. By utilizing a Pan-Tompkins based QRS detection algorithm [90], using Kubios, the R peaks for any raw ECG signal are detected automatically. In our experiment, only the BITalino (r)evolution had raw ECG outputs. This device was set to record the raw ECG at a sampling rate of 1000 Hz. This sampling frequency is adequate for HRV analysis. Nonetheless, for enhanced detection accuracy, the QRS detection algorithm employed in Kubios HRV interpolates the R peaks at 2000 Hz. This approach further improves the temporal resolution when detecting R peaks in the case of devices with lower ECG sampling rates.

Kubios performs a matched filtering technique to detect the pulse waves from the PPG raw signals. For predicting the initial pulse position, a maximum of first derivatives is used. The steepest part of the pulse corresponds to this first derivative. In the next step, a matched filter is obtained by using the correlation of the first pulse found in the previous step as the template to detect the presence and locations of a similar template (pulses) in the upcoming parts of the signal. The amount of acceptable normalized error value between the varying PPG signals and the template can be set in software preferences. The smaller the acceptance threshold is, the more similar the pulse wave has to be with the template in order to be accepted. In this study, we used Kubios to extract IBI values from raw PPG data recorded by the Empatica E4 wristband. For the rest of the devices, the RR values are automatically processed inside the device and can be downloaded as output.

2) ARTIFACT REMOVAL

Before conducting the analysis on the collected data, it needs to be preprocessed. One of the most critical steps of the preprocessing stage is artifact removal. Like any other signal, biomedical and psychophysiological signals are also prone to noises and artifacts. Common types of artifacts in HRV analysis are mostly the results of low data quality caused by factors like movement and technical errors. These noises can arise from the users' involuntary movements or environmental factors, and depending on the sensor types and their attachment location, various sensors can be affected differently. Both causes can have more negative influences, while the heart rate data are being collected in real-life and ambulatory settings. For example, PPG sensors are extremely susceptible to noise, and their signal-to-noise ratio can be adversely affected by multiple factors. A low signal-to-noise ratio, which will eventually lead to incorrect readings, may be caused by low perfusion resulting in a weak pulsatile signal and high noise coming from multiple factors such as motion or electromagnetic interference. This is why reliable artifact detection and removal methods must be employed for PPG signals, recorded with a wristwatch. Even the slight harmful interference coming from ectopic beats, which can

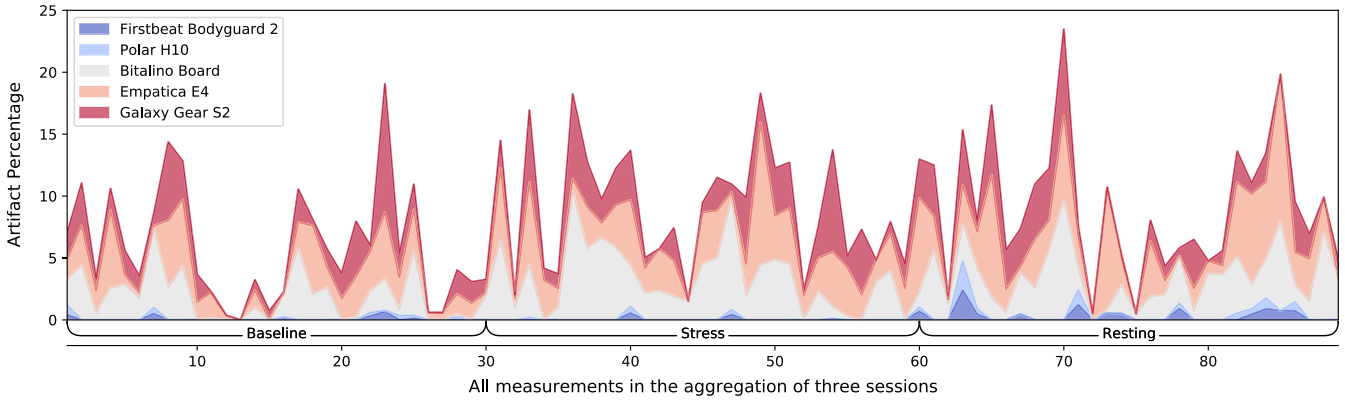


FIGURE 6. The amount of artifacts in each device during three consecutive sessions detected by automatic correction method.

cause significantly shorter inter-beat intervals (IBI) or other types of arrhythmias, are very important and must be dealt with in the most appropriate manner.

As recommended by the Taskforce [38], all artifacts in the RR time series must be either corrected or removed from analysis to reduce the chances of substantial deformities that these artifacts can cause in the HRV analysis. We applied two different types of artifact correction algorithms on the data from our five heart monitor devices. The first method, also called the automatic correction in the Kubios HRV tool, detects the artifacts from a time series proposed in [91]. The proposed method is based on time-varying thresholds calculated from the series of consecutive RR-interval variations paired with a novel classification method. This method detects extra and missed beats with a 100% true positive rate, and for the ventricular and atrial ectopic beats, true and false positive rates are 96.96% and 99.94%, respectively. The delivered sensitivity is better, and the specificity is comparable in comparison with the state-of-the-art artifact detection. The second artifact correction method we applied to our data is a threshold-based correction method, during which the ectopic beats and artifacts are corrected by comparing every RR interval against a local mean interval. In order to prevent the effects of the outliers on the RR time series, the local mean is obtained by applying median filtering on the RR time series. An interval is marked as an artifact that needs to be corrected when the difference between an RR and the local mean varies by more than a given threshold. In the Kubios HRV tool, these thresholds are named Very Low, Low, Medium, Strong and Very strong for the 0.45, 0.35, 0.25, 0.15, and 0.05 (all in seconds) threshold values, respectively. By visual inspection and manual comparison of the signal patterns before and after artifact correction, we came to the conclusion that for our reference device, Firstbeat Bodyguard 2 and Polar H10, which are both ECG devices, automatic artifact correction would be sufficient. On the other hand, for other devices, we applied Medium and Strong threshold-based artifact corrections.

As shown in Figure 6, a stacked area chart is used for better visualization of artifact values from five different devices worn simultaneously by each of 32 participants during three consecutive sessions. The x-axis represents the number of subjects. Since all three sessions are aggregated together, the lowest and highest values of x correspond to participants 1 in the Baseline session and participant 32 in the Resting session. The y-axis is in percent, representing the amount of artifacts corrected by the automatic artifact correction method for each participant recorded by each device. Artifact values for all of the devices during the aggregation of three sessions, which correspond to 25 minutes of recording, show that two ECG devices, Firstbeat Bodyguard 2 and Polar H10, have the lowest amount of artifact among all devices.

While Figure 6 depicts the amount of artifacts corrected using the automatic correction method, Figure 7 represents the percentage of beats corrected by utilizing all types of artifact correction methods. For this, we applied an unsu-



FIGURE 7. K-means clustering on the aggregation of all artifact to represent the artifact correction values in 32 clusters, Hierarchical clustering groups the devices by artifact type (Color intensity shows the percentage of corrected artifacts).

pervised clustering algorithm, K-Means, on the percentage of beats corrected in the aggregation of three sessions in each row. The number of clusters was set to be equal to the number of subjects (32) in order to represent the aggregated sessions as summaries for all subjects. Then hierarchical clustering was applied on rows. Interestingly, as a result of hierarchical clustering, the pattern of similarity between device types and the type of artifact correction modes applied to them are clearly visible in clusters. Only the Polar H10 and the reference device stand in the same cluster, and the rest of the devices are first grouped with a version of themselves that have received a higher artifact correction threshold or with other similar families of devices.

In addition to artifact detection algorithms, we also screened the data from all of the devices manually. For this, signals recorded from each participant were visually inspected and compared to both the reference and the second most accurate (Polar H10) devices to see whether there are any noisy data segments not detected by the artifact detection algorithms. In the next section, we perform the correlation comparison study of different devices in each session. In addition, the effects of different values of artifact correction thresholds on the correlation values of results obtained from different devices will be presented.

B. CORRELATION ANALYSIS

Correlation analysis generally indicates the degree to which a pair of quantitative variables are linearly related and is strictly linked to the linear regression analysis [92].

In all correlation coefficients, it is presumed that the strength of the correlation is a value between -1 and 1, indicating the highest negative correlation to the highest positive correlation, respectively, while 0 means there is no agreement at all. The most popular, best-known, and most regularly used type of correlation coefficient approach for quantitative variables, such as a measurement, is the Pearson correlation coefficient, which is the measure of the direction and strength of the linear relationship between two variables. In this approach, the correlation coefficient is calculated as the covariance of the variables divided by the product of their standard deviations. In method comparison studies, the Pearson correlation coefficient is one of the most used and first tests performed to determine whether two methods produce similar results. Although its use is still a matter of debate, most of the studies for validation of heart rate and HRV monitors and PPG based sensors have demonstrated Pearson's correlation coefficient results in their studies. For example, in a study to validate the accuracy of the Empatica E4 wrist-worn wearable sensor in different conditions, Menghini *et al.* have used Pearson product-moment correlation (r) to analyze the strength of the linear association between the measurements from Empatica E4 and an ECG device [22]. Vescio *et al.* used it to compare HRV measurement between the Kyto earlobe PPG sensor and an eMotion Faros ECG device [18]. In [23], Barrios *et al.* use Pearson's correlation as a part of their analysis to evaluate the accuracy of PPG sensors for HR and HRV measurements in-the-wild. Last but not least, in [93], Lier *et al.* have conducted a comprehensive study on validity assessment protocols for physiological signals from wearable technology. They suggested using cross-correlation as a generalization of Pearson's correlation for validity assessment of PPG and ECG devices at the signal level.

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

When the samples are not normally distributed, or when the data has strong outliers or the relationship between the variables is not linear, it is recommended to use the Spearman rank correlation method, which makes no assumptions about the distribution of the data. It evaluates monotonous relationships between two variables, whether it is linear or not, and it is equivalent to the Pearson correlation between the rank values of those two variables. In [94], Bulte *et al.* analyzed the association between HRV and PRV using Spearman's rank correlation coefficient for assessing the level of agreement between heart rate variability and pulse rate variability. Gilgen-Ammann *et al.* used Spearman to assess the correlations between the RR values from Polar H10 and an ECG holter [34] and Schrödl *et al.* utilized it to analyze the correlations between HRV features recorded by earlobe PPG and chest ECG [95]. There are also research cases in which both methods have been used [96].

In our study, we applied both methods on the time domain and frequency domain features of the HRV values recorded from five heart monitoring devices, namely, three ECG (Firstbeat Bodyguard 2, Polar H10, and BITalino) and two PPG sensors (Empatica E4 and Samsung Gear S2). Correlation analysis was performed on the aggregation of sessions to reduce the susceptibility of Pearson's correlation to the presence of strong outliers in small samples. Analysis of the correlation between the reference device and five other devices was performed with three different thresholds of the artifact correction on the devices with a higher amount of noisy data. Results of the correlation analysis for all features under study are shown in Table 3. The averages of corrected artifacts of all participants in the aggregation of three consecutive sessions expressed as the percentage of corrected beats are also presented in Table 3. When the artifact correction was set to automatic, Polar H10 showed the highest possible correlation for all features with r values of 1 in the time domain and r values higher than 0.95 for the frequency domain features. In the rest of the devices, although there were high correlations for MeanRR and pNN50, correlation

TABLE 3. Correlation values and the effects of different artifact removal thresholds.

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

values dropped dramatically in other features, mainly in the frequency domain.

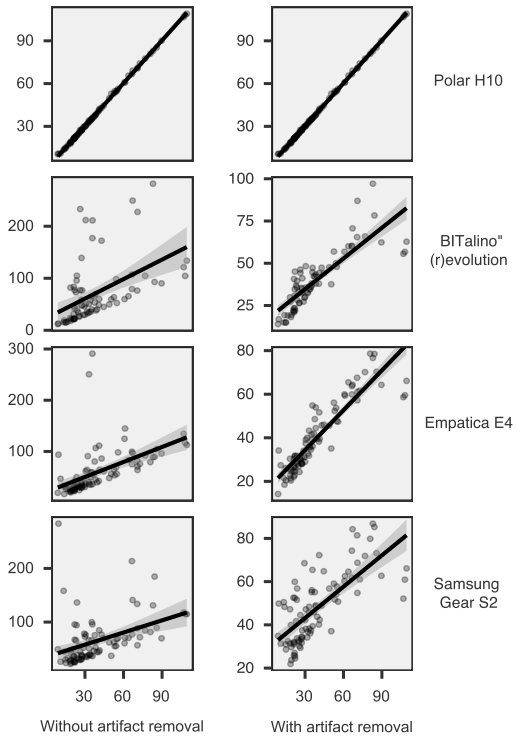
These results proved the need for further artifact corrections. In ''Medium artifact Correction'', we could see meaningful rises in all *r* values. As an instance, compared to the ''Automatic'' results, Pearson's *r* value became more than double for the Empatica E4 showing a high correlation in all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

Figure 8 shows scatter plots with regression lines (lines of best fits) and the standard error. This line minimizes the squared difference between the line of best fit and each data point. The X-axis represents the reference device (Firstbeat Bodyguard 2), and the Y-axis represents the device being compared to the reference. These plots are similar to Pearson's correlation results, which shows a positive and linear relationship between two variables. The strongest possible linear relationship occurs when the regression line's slope equals 1, and this only happens when the line of best fit lies at a 45-degree angle. By visual inspection of the plots, it is clearly visible that only Polar H10 shows a correlation almost equal to 1 in all feature types. In other devices, the regression line moves towards 1, and error rates decrease after the application of artifact removal.

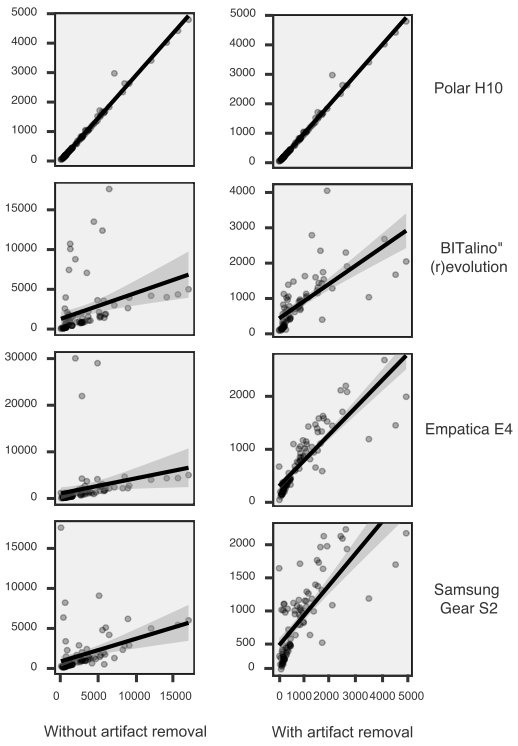
C. BLAND-ALTMAN ANALYSIS

A clinical measurement device, in our case, heart rate monitors, must have sufficient agreement with the reference golden standard devices in the quality of measurement. Correlation and regression analysis performed in the previous section are among the most popular and widely used methods in the literature. However, there is a debate that the correlation analysis only evaluates the relationship between one variable and another, not their differences, and it is not sufficient for thoroughly assessing the comparability between two different devices [97].

In a series of articles by J. M. Bland and D. G. Altman, an alternative analysis was proposed based on the quantification of the agreement between two quantitative measurements by studying the mean difference and constructing the limits of agreement [98]–[101]. The Bland-Altman plot and analysis, which is also called the limits of agreement or the mean-difference, is a comparison technique used for making a comparison between two measurements of the same variable. For instance, an obtrusive and expensive heart rate measurement system might be compared with a low-cost and unobtrusive one. The Bland-Altman (BA) method is a graphical approach in which the discrepancies between the two methods are plotted against their means [99]. Any systematic variations between the two pairwise measurements, such as fixed or proportional bias and the intensity of possible outliers, can be identified using Bland-Altman plots. While y-values depict the differences between the two measurements, the pairwise measurements' averages are assigned to the x-axis. There also other horizontal lines parallel to the x-axis, representing the average of differences (estimated bias) and the limits of agreement. Limits of agreement (LoA) in a Bland-Altman plot are the averages of the differences



(a) RMSSD (ms)



(b) HF (ms^2)

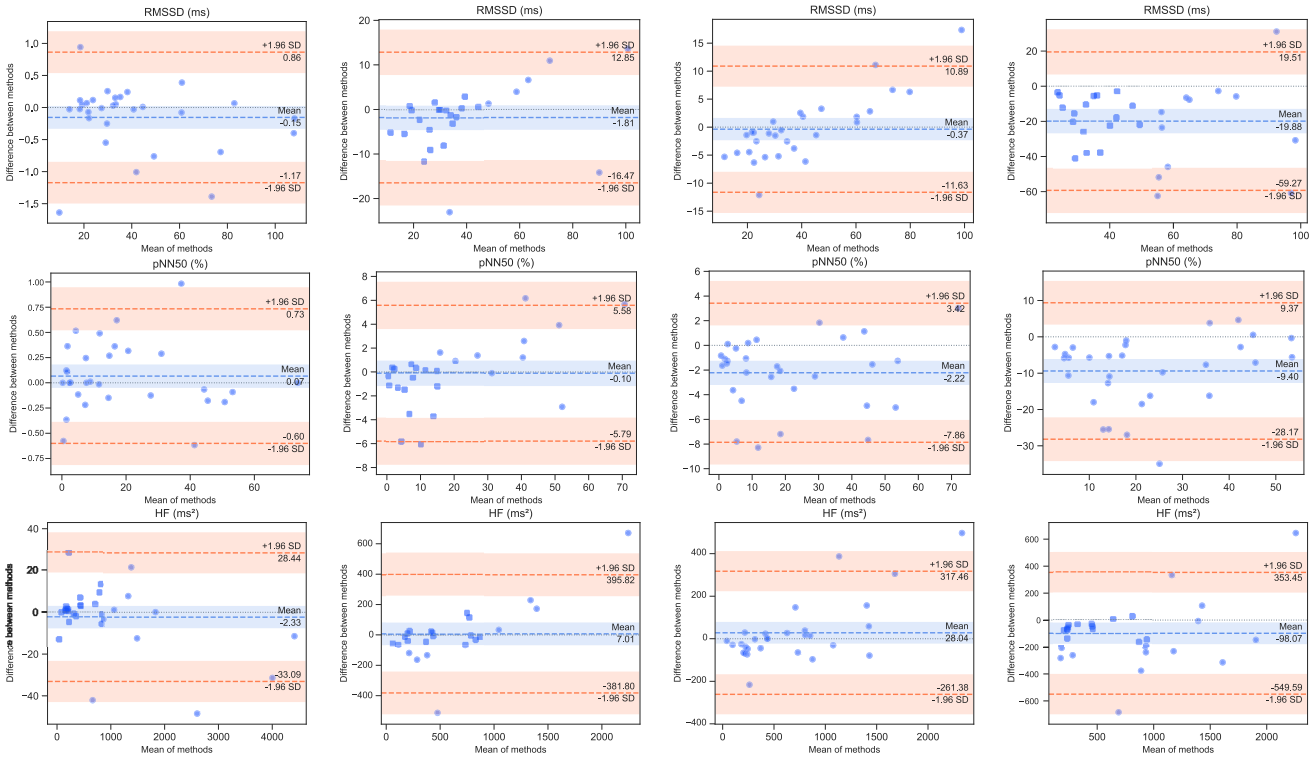
FIGURE 8. Scatter plots with linear regression line and standard error depicting the effects of artifact removal on the increase in correlation values.

± 1.96 times its standard deviation. Computing the 95% limits of agreement (LoA) helps us understand how far apart the HRV feature measurements recorded by two different devices are more likely to be for most subjects. The Bland-Altman only depicts the ranges between the upper and lower limits of the agreement without asserting whether these limits should be considered acceptable. If a predefined a priori exists as the acceptable limits, it should be verified whether the limits of agreement on the Bland-Altman plot exceed this allowed difference or not. If the LoA are within the acceptable limits, we can state that the values inside the mean ± 1.96 standard deviation of the differences are not clinically significant, and the two measurement techniques (devices) are in agreement and can be used interchangeably.

To the best of our knowledge, there is no a priori acceptable limit defined for HRV features hitherto. Some researchers have proposed their own methods, such as accepting $\pm 50\%$ of the mean of the reference value as an acceptable limit or creating acceptable ratios by calculating the ratio of half the range of limits of agreement and the mean of averages. We believe that acceptable limits of accuracy and acceptability for HRV measurements must be issued from professional organizations specialized in this field. For example, the American National Standard for the advancement of medical instrumentation (ANSI/AAMI) accepts heart rate monitors as accurate if their measuring error does not exceed ± 5 bpm or $\pm 10\%$ [102].

Accordingly, we interpret our Bland-Altman results based on visual inspection of the relevant plots and compare our devices against each other and to the reference device. Bland-Altman plots are represented in Figures 9, 10, and 11. These plots represent six HRV features in three different sessions. The artifact correction levels for the Bland-Altman analysis is set to Automatic for the reference device and Polar H10. For the rest of the devices, this value is set to Strong. Quantitative results of the BA plots with two more features (eight in total) are depicted in Tables 4, 5, and 6. In most of our Bland-Altman plots, the scattered points' density is higher on the left side of the plots. Such a problem can easily be solved by performing a log transformation on the samples. However, we can reveal any possible relationship between the measurement differences of both methods and the magnitude of measurements by plotting the actual values of the HRV features. Except for the Polar H10 and the reference device, a maximum of two strong outliers were removed from the rest of the devices in order to prevent the serious negative effects of influential outliers on BA results. Since the number of devices in this study is four (five, including the Firstbeat Bodyguard 2 as the reference device), instead of showing all 96 BA plots, we decided to bring only the most important ones. However, detailed statistics of all 96 cases are presented in Tables 4, 5, and 6.

In the following paragraphs, we will discuss the results of the Bland-Altman plots based on visual inspection and comparison between different devices in each scenario. For all features throughout three sessions, Polar H10 chest strap shows the largest agreement with the reference device, and its mean bias in all features is near zero. We expected to achieve the same amount of high-quality data based on



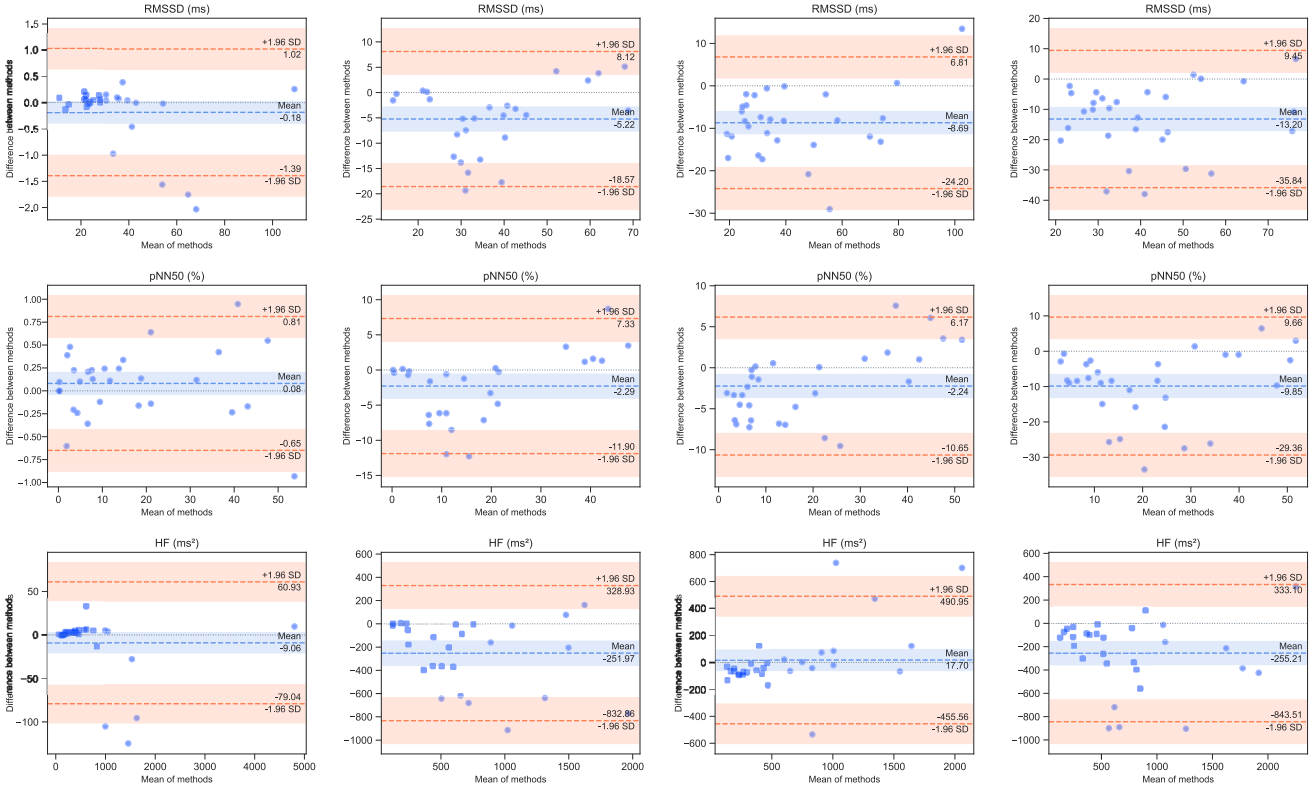
(a) Polar H10 (b) BITalino (r)evolution (c) Empatica E4 (d) Samsung Gear S2

FIGURE 9. **Bland-Altman plots for time domain and frequency domain features - Baseline session.**

TABLE 4. **Bland-Altman results for the "Baseline" session.**

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

previous experiences we had during tests and data acquisition using BITalino (r)evolution kit, which performed very weakly despite high expectations. Since BITalino (r)evolution is a board kit, it is highly prone to get contaminated with a high amount of artifacts due to subject movements. Misplaced sensor placement by subjects can also cause poor skin contact and attenuated signal reading capabilities since the device is not wearable. This happened in our study, as well. BITalino (r)evolution suffered from lots of artifacts, mostly because of sensor misplacement and other reasons mentioned earlier. However, by checking the correlation and BA results for BITalino (r)evolution, we can witness the effectiveness of proper artifact removal on its noisy data. Except for the Stress session in which BITalino (r)evolution shows clear signs of systematic errors by producing mean shifts way below or above zero, in the time-domain features of the Baseline



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

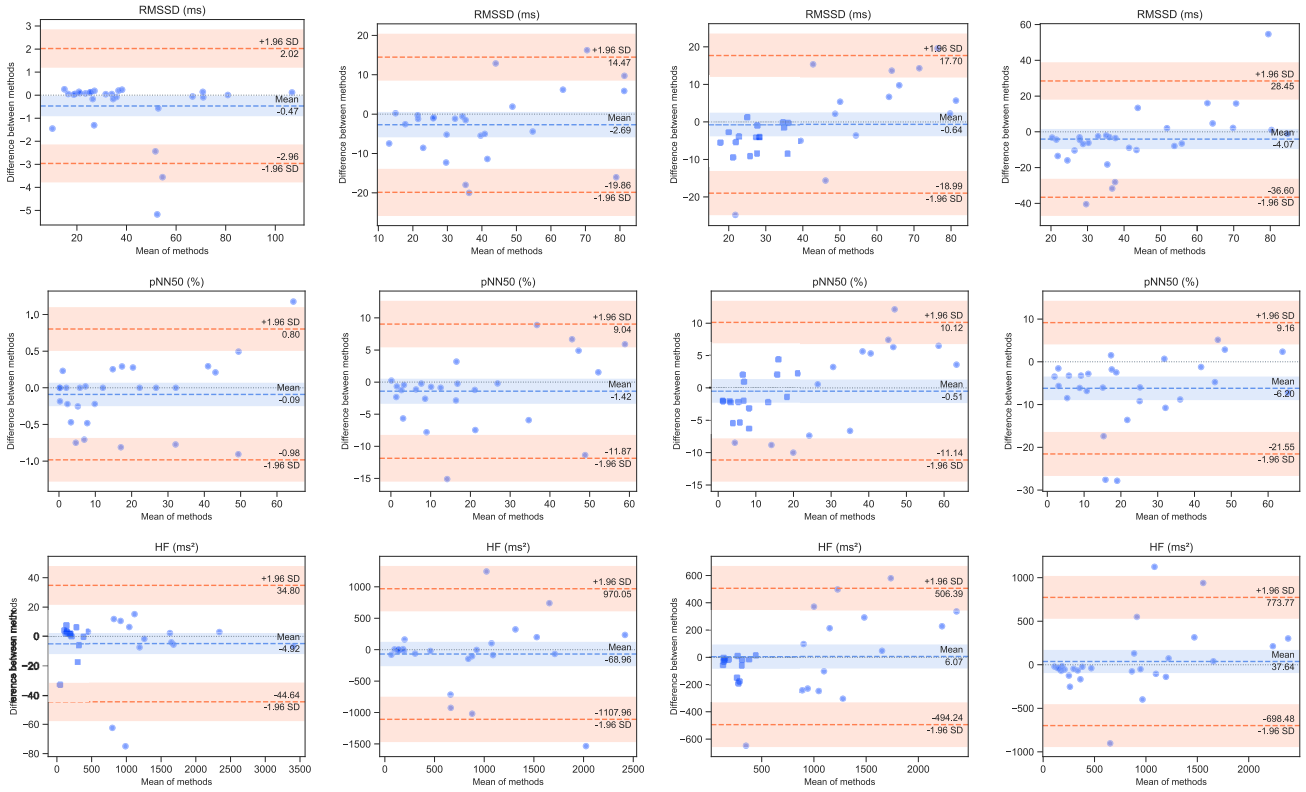
FIGURE 10. Bland-Altman plots for time domain and frequency domain features - Stress session.

TABLE 5. Bland-Altman results for the "Stress" session.

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

and Recovery sessions, it performs quite better compared to the Samsung Gear S2 PPG wristband, especially in the normalized HF and LF features in Recovery session. BITalino (r)evolution's abysmal performance in Stress session matches its higher artifact values in that session, as shown in Figure 6.

Although the Empatica E4 wristband shows signs of proportional error in RMSSD only in the Baseline session, this trend does not seem to happen again in two subsequent sessions. While both PPG devices show errors in the pNN50 time-domain feature, Empatica E4 shows good agreement in the Baseline session. Similar to BITalino (r)evolution,



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

FIGURE 11. Bland-Altman plots for time domain and frequency domain features - Resting session.

TABLE 6. Bland-Altman results for the "Resting" session.

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

all other devices except Polar H10 show systematic errors in the stress session's time-domain features. For the rest of the features in the Stress and the Resting session, Empatica E4 shows overall good performance with slight cases of proportional error in time-domain features.

Samsung Gear S2 offers the weakest performance among all devices under study. It shows systematic errors nearly in all time-domain features. Since all error patterns in time-domain features are similar, it can be adjusted for. However, a high number of errors are also present in the frequency domain, making this device the least reliable one, with the lowest amount of agreement with our reference device.

In some of the plots, we can observe that samples tend to tilt towards limits of agreement, mostly at the right side of

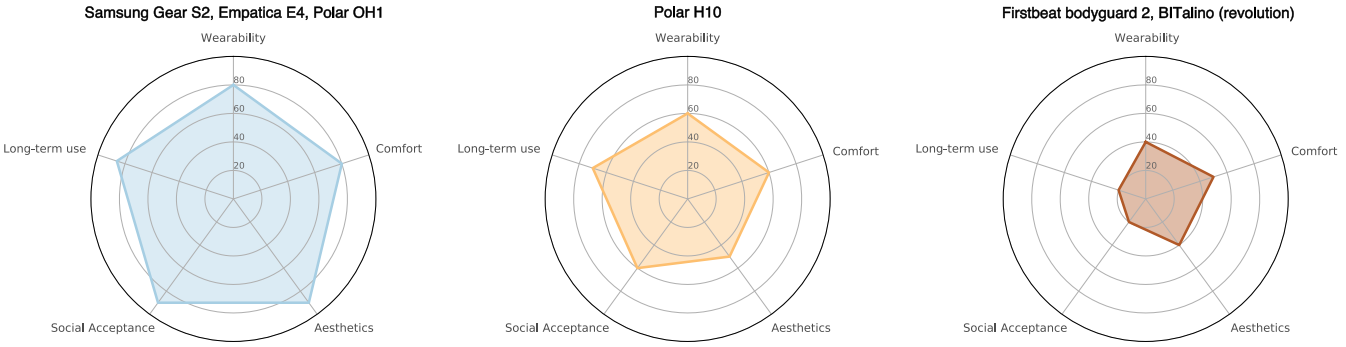


FIGURE 12. Comparision of wearability, comfort, aesthetics, social acceptance and long-term use based on user feedback.

the plot. In such a case, we can deduce that the variation of measurement on that particular feature strongly depends on the magnitude of measurements. Besides that, some of the proportional errors are also caused by the concentration of samples on the left side of the plot. These problems can be eliminated either by log transformation or by reducing the windows size, which will result in an increased sample size. It should also be noted that a consistent measurement bias can be corrected by subtracting the mean difference from the second method.

D. USERS' VIEWS AND EXPERIENCES: WEARABILITY, COMFORT, AESTHETICS, SOCIAL ACCEPTANCE, AND LONG-TERM USE

In this section, we present details on interviews conducted at the last stage of the study. In particular, we present participants' experiences on the sensors' wearability and comfort, their views on its aesthetics, social acceptance, and long-term use. The interview data were transcribed for thematic analysis [62] using Atlas.ti [64]. Atlas.ti is a software tool for qualitative analysis of textual data. It allows adding documents and marking quotes from the text, and transforming them into codes that are later used in theme development. We generated over 100 codes involving participants' interpretations and experiences with the sensors, which we later used to develop themes detailed below.

Participants described wrist-worn devices such as Galaxy Gear 2, Empatica E4, and also the Polar OH1 armband as more wearable compared to other sensors: *"I think the ones on the wrist are easier to wear on for everyday life"* - P14 (Participant 14). They described wrist-worn devices as comfortable and lightweight: *"These devices were really comfortable"* - P3. Similarly, Polar OH1 was also considered to be quite comfortable, and all participants mentioned that they did not notice wearing one during the data collection process until they had to take it off: *"I did not even feel this Polar OH1, that is probably the most comfortable and it is easier to put on and off. You just have to slide your arm straight in"*. However, some participants felt Empatica E4 to be a little heavier and uncomfortable than other wrist-worn devices because of its electrodes constantly pressing against the skin: *"Empatica E4 felt quite sore and heavy after using it for a while"* - P16. Compared to the wrist and arm-worn devices, participants found chest strap, i.e., Polar H10, to be challenging to wear on as it requires the bands to be wrapped around the chest tightly. Moreover, they mentioned that the bands put pressure on the abdomen, which felt a little uncomfortable while sitting: *"Chest strap is not that much tight, but they can restrict your breathing when you are sitting"* - P5. Bodyguard 2 and BITalino (r)evolution require two or three Ag/AgCl (silver/silver-chloride) electrodes, respectively. These electrodes are self-adhesive and should be placed on particular positions on the skin for precise signal acquisition. All participants found these electrodes to be uncomfortable while taking off the skin: *"When you take it off the electrodes, it hurts"* - P15. Participants also describe these devices as not quite convenient for daily lives because it requires multiple wires: *"For devices with electrodes, I could feel something hanging from my body"* - P21.

In terms of design and aesthetics, participants preferred the Samsung Gear S2 smartwatch over Empatica E4. They mentioned that besides offering sensing capabilities, smartwatches also provide other functionalities for everyday use. They described it as *"appealing"*, *"everyday objects"*, and *"stylish"* - P7, P18, P31. Participants also mentioned that smartwatches are more socially acceptable than other sensing devices, and they are more likely to wear them over the long term. Chest strap, on the other hand, were hidden under the clothes and therefore were not thought to attract any social judgment from other people, but participants associated them as being sporty: *"They are not as comfortable, and you might only wear them for specific things like training or to gauge fitness levels"* - P10. Furthermore, participants regarded First beat bodyguard 2 and BITalino (r)evolution as medical devices. They mentioned that they would not like to wear it in public places and over longer times as they are obtrusive and would attract attention and lead other people to believe that the wearer has a medical condition: *"I imagine the [First beat bodyguard 2 and BITalino (r)evolution] look like people wear in the hospital [and] would lead people to assume that you are wearing for medical purposes rather than recreational"* - P23.

At the end of the interview, participants filled out a 5-point Likert scale questionnaire on experiences discussed above, the results of which are summarized in Figure 12. Results mirror the qualitative analysis highlighting that participants' preference of wrist and arm-worn devices in terms of aesthetics, wearability and comfort, long-term use, and social acceptance, followed by Polar H10 and First beat bodyguard 2 and BITalino (r)evolution.

V. CONCLUSION

We chose six of the most common biosensors for heart rate monitoring during Baseline, Stress, and Relaxation sessions in a controlled laboratory setting. We performed quantitative analysis on HRV data and accessed the reliability and accuracy of five of these devices. Our results highlight that in all sessions, Polar H10 achieved the highest correlation and agreement levels with the reference device (First beat bodyguard 2), and also had the lowest amount of artifacts, followed by Empatica E4, BITalino (r)evolution and Samsung Gear S2. Wrist-worn devices (Empatica E4, Samsung Gear S2) showed signs of systematic and slight proportional errors in the agreement analysis and much lower correlation values than the reference device and the Polar H10, especially in the frequency domain features. We expected to witness low amounts of artifacts in all devices since the data acquisition was performed in a controlled lab environment. In addition, participants were asked to wear all the devices simultaneously, meaning that the possibility of any noise contamination caused either by movements or by environmental factors was almost equal for all devices. However, Empatica E4, Samsung Gear S2, and BITalino (r)evolution suffered from many artifacts. The effects of these artifacts on the correlation and agreement analysis were studied, and it is concluded that proper artifact removal thresholds can decrease the adverse effects of the artifacts to the minimum. We also conducted interviews with participants and analyzed their views and experiences on user acceptance of the sensing devices using thematic analysis. Our findings suggest that participants prefer Samsung Gear S2, Empatica E4 and Polar OH1 in terms of aesthetics, wearability, and comfort, followed by Polar H10, First beat bodyguard 2, and BITalino (r)evolution. Moreover, participants mentioned that First beat bodyguard 2, BITalino (r)evolution followed by Polar H10 are more likely to invite social judgment from others, and they would not want to wear it in public. Participants preferred Polar H10 for short-term use, and Samsung Gear S2, Empatica E4 and Polar OH1 over long-time.

Increasing work on HRV assessment has used a variety of mobile, mostly wearable inconspicuous sensing devices. These devices use electrocardiography (ECG) and photoplethysmography (PPG) to capture heart data and employ lightweight technologies that can be worn easily and unobtrusively. Our work contributes to the existing literature by introducing a mixed-methods approach comprising both quantitative and qualitative data analysis techniques across six heart rate monitoring devices on key body locations. Choosing the appropriate device must be decided based on the type of activity and predefined goals. The quantitative data analysis presented in our paper would help guide researchers of affective biofeedback technologies for stress to carefully chose the appropriate sensing device while considering usability and user acceptance of the device. For applications with low movements and physical activity where there is no need for near-absolute accuracy, wrist-worn devices i.e. Empatica E4 and Samsung Gear S2 smartwatch offer a trade-off between acceptable accuracy and long-term wearability, and comfort. For near medical-grade accuracy, we advocate for Polar H10 and Firstbeat bodyguard 2. However, they provide limited wearability and comfort and are not suitable for wearing over extended periods of time. BITalino offers the most customizable settings and usage with a variety of sensors, and it is suitable for exploring multi-modal biosignal analysis. However, the lack of wearability makes it more suitable to explore in controlled settings with limited movement.

Certain constraints existing in some devices can lead to shortcomings in scenarios that require longer HRV recordings. For instance, the limited battery life in Samsung Gear S2, and BITalino (r)evolution kit, makes it impractical to conduct continuous recordings of more than three-four hours. Another limitation is the need for constant Bluetooth connectivity to a third-party mobile or computer application for data acquisition in BITalino (r)evolution kit, making such devices less suitable for recordings outside the lab environment.

Our results are obtained in laboratory conditions, in the same 70-minute long sessions while the subjects are sitting and are exposed to the identical baseline, stress, and relaxation protocols. We argue that researchers and practitioners should be aware of the strengths and limitations of the consumer heart rate devices when conducting studies. We hope that our work would be beneficial in guiding users in exploring the trade-off between data accuracy and usability when choosing a heart rate and HRV monitoring device.

ACKNOWLEDGMENT

(Muhammad Umair and Niaz Chalabianloo contributed equally to this work.)

REFERENCES

[1] G. Fink, "Stress: Definition and history," in *Encyclopedia of Neuroscience*, vol. 9, L. R. Squire, Ed. Oxford, U.K.: Academic, 2009, pp. 549–555.

[2] T. G. Pickering, "Mental stress as a causal factor in the development of hypertension and cardiovascular disease," *Current Hypertension Rep.*, vol. 3, no. 3, pp. 249–254, Jun. 2001.

[3] N. Schneiderman, G. Ironson, and S. D. Siegel, "Stress and health: Psychological, behavioral, and biological determinants," *Annu. Rev. Clin. Psychol.*, vol. 1, no. 1, pp. 607–628, Apr. 2005.

[4] *Work-Related Stress, Anxiety or Depression Statistics in Great Britain, 2019*, Health Saf. Executive, Merseyside, U.K., 2019.

[5] J. Hassard, K. Teoh, T. Cox, P. Dewe, M. Cosmar, R. Grundler, D. Flemming, B. Cosemans, and K. Van den Broek, "Calculating the cost of work-related stress and psychosocial risks," Publications Office Eur. Union, Luxembourg City, Luxembourg, Tech. Rep., 2014.

[6] B. Yu, M. Funk, J. Hu, Q. Wang, and L. Feijs, "Biofeedback for everyday stress management: A systematic review," *Frontiers ICT*, vol. 5, p. 23, Sep. 2018.
[7] J. B. Lemaire, J. E. Wallace, A. M. Lewin, J. de Grood, and J. P. Schaefer, "The effect of a biofeedback-based stress management tool on physician stress: A randomized controlled clinical trial," *Open Med.*, vol. 5, no. 4, p. e154, 2011.
[8] A. Dillon, M. Kelly, I. H. Robertson, and D. A. Robertson, "Smartphone applications utilizing biofeedback can aid stress reduction," *Frontiers Psychol.*, vol. 7, p. 832, Jun. 2016.
[9] H. Al Osman, H. Dong, and A. El Saddik, "Ubiquitous biofeedback serious game for stress management," *IEEE Access*, vol. 4, pp. 1274–1286, 2016.
[10] B. Yu, J. Hu, M. Funk, and L. Feijs, "DeLight: Biofeedback through ambient light for stress intervention and relaxation assistance," *Pers. Ubiquitous Comput.*, vol. 22, no. 4, pp. 787–805, Aug. 2018.
[11] R. Gevirtz, "The promise of heart rate variability biofeedback: Evidence-based applications," *Biofeedback*, vol. 41, no. 3, pp. 110–120, Sep. 2013, doi: 10.5298/1081-5937-41.3.01.
[12] P. L. A. Schoenberg and A. S. David, "Biofeedback for psychiatric disorders: A systematic review," *Appl. Psychophysiol. Biofeedback*, vol. 39, no. 2, pp. 109–135, Jun. 2014, doi: 10.1007/s10484-014-9246-9.
[13] V. C. Goessl, J. E. Curtiss, and S. G. Hofmann, "The effect of heart rate variability biofeedback training on stress and anxiety: A meta-analysis," *Psychol. Med.*, vol. 47, no. 15, pp. 2578–2586, Nov. 2017, doi: 10.1017/S0033291717001003.
[14] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo, "Stress and heart rate variability: A meta-analysis and review of the literature," *Psychiatry Investig.*, vol. 15, no. 3, p. 235, 2018.
[15] K. Herbell and J. A. Zauszniewski, "Reducing psychological stress in peripartum women with heart rate variability biofeedback: A systematic review," *J. Holistic Nursing*, vol. 37, no. 3, pp. 273–285, Sep. 2019.
[16] R. McCraty and F. Shaffer, "Heart rate variability: New perspectives on physiological mechanisms, assessment of self-regulatory capacity, and health risk," *Global Adv. Health Med.*, vol. 4, no. 1, pp. 46–61, Jan. 2015.
[17] Y. S. Can, N. Chalabianloo, D. Ekiz, and C. Ersoy, "Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study," *Sensors*, vol. 19, no. 8, p. 1849, Apr. 2019.
[18] B. Vescio, M. Salsone, A. Gambardella, and A. Quattrone, "Comparison between electrocardiographic and earlobe pulse photoplethysmographic detection for evaluating heart rate variability in healthy subjects in short- and long-term recordings," *Sensors*, vol. 18, no. 3, p. 844, 2018.
[19] E. Mejía-Mejía, K. Budidha, T. Y. Abay, J. M. May, and P. A. Kyriacou, "Heart rate variability (HRV) and pulse rate variability (PRV) for the assessment of autonomic responses," *Frontiers Physiol.*, vol. 11, p. 779, Jul. 2020.
[20] R. Rauh, R. Limley, R.-D. Bauer, M. Radespiel-Troger, and M. Mueck-Weymann, "Comparison of heart rate variability and pulse rate variability detected with photoplethysmography," *Proc. SPIE*, vol. 5474, pp. 115–126, Aug. 2004.
[21] C. Yu, Z. Liu, T. McKenna, A. T. Reisner, and J. Reifman, "A method for automatic identification of reliable heart rates calculated from ECG and PPG waveforms," *J. Amer. Med. Inform. Assoc.*, vol. 13, no. 3, pp. 309–320, May 2006.
[22] L. Menghini, E. Gianfranchi, N. Cellini, E. Patron, M. Tagliabue, and M. Sarlo, "Stressing the accuracy: Wrist-worn wearable sensor validation over different conditions," *Psychophysiology*, vol. 56, no. 11, p. e13441, 2019.
[23] L. Barrios, P. Oldrati, S. Santini, and A. Lutterotti, "Evaluating the accuracy of heart rate sensors based on photoplethysmography for in-the-wild analysis," in *Proc. 13th EAI Int. Conf. Pervas. Comput. Technol. Healthcare*, May 2019, pp. 251–261.
[24] E. Jo, K. Lewis, D. Directo, M. J. Kim, and B. A. Dolezal, "Validation of biofeedback wearables for photoplethysmographic heart rate tracking," *J. Sports Sci. Med.*, vol. 15, no. 3, p. 540, 2016.
[25] H. Ehmen, M. Haesner, I. Steinke, M. Dorn, M. Gövercin, and E. Steinhagen-Thiessen, "Comparison of four different mobile devices for measuring heart rate and ECG with respect to aspects of usability and acceptance by older people," *Appl. Ergonom.*, vol. 43, no. 3, pp. 582–587, May 2012, doi: 10.1016/j.apergo.2011.09.003.
[26] A. Bryman, "Integrating quantitative and qualitative research: How is it done?" *Qualitative Res.*, vol. 6, no. 1, pp. 97–113, Feb. 2006.
[27] S. Almalki, "Integrating quantitative and qualitative data in mixed methods research–challenges and benefits," *J. Educ. Learn.*, vol. 5, no. 3, pp. 288–296, 2016.
[28] T. Föhr, A. Tolvanen, T. Myllymäki, E. Järvelä-Reijonen, S. Rantala, R. Korpela, K. Peuhkuri, M. Kolehmainen, S. Puttonen, R. Lappalainen, H. Rusko, and U. M. Kujala, "Subjective stress, objective heart rate variability-based stress, and recovery on workdays among overweight and psychologically distressed individuals: A cross-sectional study," *J. Occupational Med. Toxicol.*, vol. 10, no. 1, p. 39, Dec. 2015.
[29] K. Kutt, W. Binek, P. Misiak, G. J. Nalepa, and S. Bobek, "Towards the development of sensor platform for processing physiological data from wearable sensors," in *Proc. Int. Conf. Artif. Intell. Soft Comput.*, 2018, pp. 168–178.
[30] C. Chen, C. Li, C.-W. Tsai, and X. Deng, "Evaluation of mental stress and heart rate variability derived from wrist-based photoplethysmography," in *Proc. IEEE Eurasia Conf. Biomed. Eng., Healthcare Sustainability (ECBIOS)*, May 2019, pp. 65–68.
[31] T. Bhowmik, J. Dey, and V. N. Tiwari, "A novel method for accurate estimation of HRV from smartwatch PPG signals," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 109–112.
[32] Y. S. Can, N. Chalabianloo, D. Ekiz, J. Fernandez-Alvarez, C. Repetto, G. Riva, H. Iles-Smith, and C. Ersoy, "Real-life stress level monitoring using smart bands in the light of contextual information," *IEEE Sensors J.*, vol. 20, no. 15, pp. 8721–8730, Aug. 2020.
[33] J. Parak and I. Korhonen. (2015). *Accuracy of Firstbeat Bodyguard 2 Beat-to-Beat Heart Rate Monitor*. Accessed: Jan. 17, 2021. [Online]. Available: https://www.firstbeat.com/wp-content/uploads/2015/10/white_paper_bodyguard2_final.pdf
[34] R. Gilgen-Ammann, T. Schweizer, and T. Wyss, "RR interval signal quality of a heart rate monitor and an ECG holter at rest and during exercise," *Eur. J. Appl. Physiol.*, vol. 119, no. 7, pp. 1525–1532, Jul. 2019.
[35] C. McCarthy, N. Pradhan, C. Redpath, and A. Adler, "Validation of the empatica e4 wristband," in *Proc. IEEE EMBS Int. Student Conf. (ISC)*, May 2016, pp. 1–4.
[36] M. Alfaras, W. Primett, M. Umair, C. Windlin, P. Karpashevich, N. Chalabianloo, D. Bowie, C. Sas, P. Sanches, K. Höök, C. Ersoy, and H. Gamboa, "Biosensing and actuation—Platforms coupling body input-output modalities for affective technologies," *Sensors*, vol. 20, no. 21, p. 5968, 2020.
[37] P. Sanches, A. Janson, P. Karpashevich, C. Nadal, C. Qu, C. Daudén Roquet, M. Umair, C. Windlin, G. Doherty, K. Höök, and C. Sas, "HCI and affective health: Taking stock of a decade of studies and charting future research directions," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2019, pp. 1–17.
[38] 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," *Ann. Noninvasive Electrocardiol.*, vol. 1, no. 2, pp. 151–181, Apr. 1996.
[39] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers Public Health*, vol. 5, p. 258, Sep. 2017.
[40] S. Laborde, E. Mosley, and J. F. Thayer, "Heart rate variability and cardiac vagal tone in psychophysiological research—Recommendations for experiment planning, data analysis, and data reporting," *Frontiers Psychol.*, vol. 8, p. 213, Feb. 2017.
[41] S. Akselrod, D. Gordon, F. Ubel, D. Shannon, A. Berger, and R. Cohen, "Power spectrum analysis of heart rate fluctuation: A quantitative probe of beat-to-beat cardiovascular control," *Science*, vol. 213, no. 4504, pp. 220–222, Jul. 1981, doi: 10.1126/science.6166045.
[42] R. Lane, K. Mcrae, E. Reiman, K. Chen, G. Ahern, and J. Thayer, "Neural correlates of heart rate variability during emotion," *NeuroImage*, vol. 44, no. 1, pp. 213–222, Jan. 2009, doi: 10.1016/j.neuroimage.2008.07.056.
[43] S. Boonnithi and S. Phongsuphap, "Comparison of heart rate variability measures for mental stress detection," in *Proc. Comput. Cardiol.*, Sep. 2011, pp. 85–88.
[44] D. L. Frank, L. Khorshid, J. F. Kiffer, C. S. Moravec, and M. G. McKee, "Biofeedback in medicine: Who, when, why and how?" *Ment. Health Fam. Med.*, vol. 7, no. 2, pp. 85–91, Jun. 2010.
[45] M. Umair, C. Sas, and M. H. Latif, "Towards affective chronometry: Exploring smart materials and actuators for real-time representations of changes in arousal," in *Proc. Designing Interact. Syst. Conf.*, 2019, pp. 1479–1494.

[46] D. McDuff, A. Karlson, A. Kapoor, A. Roseway, and M. Czerwinski, "AffectAura: An intelligent system for emotional memory," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2012, pp. 849–858.

[47] J. Costa, A. T. Adams, M. F. Jung, F. Guimbretière, and T. Choudhury, "EmotionCheck: Leveraging bodily signals and false feedback to regulate our emotions," in *Proc. ACM Int. Joint Conf. Pervas. Ubiquitous Comput.*, Sep. 2016, pp. 758–769.

[48] M. Umair, C. Sas, and M. Alfaras, "ThermoPixels: Toolkit for personalizing arousal-based interfaces through hybrid crafting," in *Proc. DIS*, 2020, pp. 1017–1032.

[49] A. Thieme, J. Wallace, P. Johnson, J. McCarthy, S. Lindley, P. Wright, P. Olivier, and T. D. Meyer, "Design to promote mindfulness practice and sense of self for vulnerable women in secure hospital services," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Apr. 2013, pp. 2647–2656.

[50] J. Vidyarthi, B. E. Riecke, and D. Gromala, "Sonic cradle: Designing for an immersive experience of meditation by connecting respiration to music," in *Proc. Designing Interact. Syst. Conf.*, 2012, pp. 408–417.

[51] P. M. Lehrer and R. Gevirtz, "Heart rate variability biofeedback: How and why does it work?" *Frontiers Psychol.*, vol. 5, p. 756, Jul. 2014.

[52] M. Umair and C. Sas, "Problems in practice: Critiquing on thermochromic based displays for representing skin conductance," in *Proc. CHI*, 2019, pp. 1–5.

[53] A. Ghandeharioun and R. Picard, "BrightBeat: Effortlessly influencing breathing for cultivating calmness and focus," in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst.*, 2017, pp. 1624–1631.

[54] B. B. Brown, *Stress and the Art of Biofeedback*. Oxford, U.K.: Harper & Row, 1977.

[55] A. Schäfer and J. Vagedes, "How accurate is pulse rate variability as an estimate of heart rate variability: A review on studies comparing photoplethysmographic technology with an electrocardiogram," *Int. J. Cardiol.*, vol. 166, no. 1, pp. 15–29, Jun. 2013.

[56] E. Gil, M. Orini, R. Bailon, J. M. Vergara, L. Mainardi, and P. Laguna, "Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions," *Physiol. Meas.*, vol. 31, no. 9, p. 1271, 2010.

[57] W. H. Lin, D. Wu, C. Li, H. Zhang, and Y. T. Zhang, "Comparison of heart rate variability from PPG with that from ECG," in *Proc. IFMBE*, vol. 42, 2014, pp. 213–215.

[58] O. Binsch, T. Wabeke, and P. Valk, "Comparison of three different physiological wristband sensor systems and their applicability for resilience- and work load monitoring," in *Proc. 13th Annu. Body Sensor Netw. Conf. (BSN)*, Jun. 2016, pp. 272–276.

[59] Z. Ge, P. W. C. Prasad, N. Costadopoulos, A. Alsadoon, A. K. Singh, and A. Elchouemi, "Evaluating the accuracy of wearable heart rate monitors," in *Proc. ICACCA*, Nov. 2016, pp. 1–6.

[60] S. Ollander, C. Godin, A. Campagne, and S. Charbonnier, "A comparison of wearable and stationary sensors for stress detection," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2016, pp. 4362–4366.

[61] P. Bazeley, *Qualitative Data Analysis: Practical Strategies*. Newbury Park, CA, USA: Sage, 2013.

[62] V. Braun and V. Clarke, "Thematic analysis," in *APA Handbook of Research Methods in Psychology: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological*, vol. 2. Washington, DC, USA: American Psychological Association, 2012, pp. 57–71.

[63] J. Fereday and E. Muir-Cochrane, "Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development," *Int. J. Qualitative Methods*, vol. 5, no. 1, pp. 80–92, Mar. 2006.

[64] S. Friese, *Qualitative Data Analysis With ATLAS. Ti*. Newbury Park, CA, USA: Sage, 2019.

[65] P. Bazeley and K. Jackson, *Qualitative Data Analysis With NVivo*. Newbury Park, CA, USA: Sage, 2013.

[66] A. Ståhl, K. Höök, M. Svensson, A. S. Taylor, and M. Combetto, "Experiencing the affective diary," *Pers. Ubiquitous Comput.*, vol. 13, no. 5, pp. 365–378, Jun. 2008, doi: 10.1007/s00779-008-0202-7.

[67] P. Sanches, K. Höök, E. Vaara, C. Weymann, M. Bylund, P. Ferreira, N. Peira, and M. Sjölinder, "Mind the body: Designing a mobile stress management application encouraging personal reflection," in *Proc. 8th ACM Conf. Designing Interact. Syst. (DIS)*, 2010, pp. 47–56, doi: 10.1145/1858171.1858182.

[68] P. Ferreira, P. Sanches, K. Höök, and T. Jaensson, "License to chill: How to empower users to cope with stress," in *Proc. 5th Nordic Conf. Hum.-Comput. Interact. Building Bridges (NordiCHI)*, 2008, pp. 123–132, doi: 10.1145/1463160.1463174.

[69] M. Simonnet and B. Gourvennec, "Heart rate sensors acceptability: Data reliability vs. Ease of use," in *Proc. IEEE 13th Int. Conf. Wearable Implant. Body Sensor Netw. (BSN)*, Jun. 2016, pp. 94–98.

[70] G. Regalia, F. Onorati, M. Lai, C. Caborni, and R. W. Picard, "Multimodal wrist-worn devices for seizure detection and advancing research: Focus on the empatica wristbands," *Epilepsy Res.*, vol. 153, pp. 79–82, Jul. 2019.

[71] Y. S. Can, N. Chalabianloo, D. Ekiz, J. Fernandez-Alvarez, G. Riva, and C. Ersoy, "Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches," *IEEE Access*, vol. 8, pp. 38146–38163, 2020, doi: 10.1109/ACCESS.2020.2975351.

[72] E. Weston, P. Le, and W. S. Marras, "A biomechanical and physiological study of office seat and tablet device interaction," *Appl. Ergonom.*, vol. 62, pp. 83–93, Jul. 2017.

[73] C. Ottaviani, L. Shahabi, M. Tarvainen, I. Cook, M. Abrams, and D. Shapiro, "Cognitive, behavioral, and autonomic correlates of mind wandering and perseverative cognition in major depression," *Frontiers Neurosci.*, vol. 8, p. 433, Jan. 2015.

[74] N. Ahmed and Y. Zhu, "Early detection of atrial fibrillation based on ECG signals," *Bioengineering*, vol. 7, no. 1, p. 16, Feb. 2020.

[75] D. Batista, H. Plácido da Silva, A. Fred, C. Moreira, M. Reis, and H. A. Ferreira, "Benchmarking of the BITalino biomedical toolkit against an established gold standard," *Healthcare Technol. Lett.*, vol. 6, no. 2, pp. 32–36, Apr. 2019.

[76] K. E. Speer, S. Semple, N. Naumovski, and A. J. McKune, "Measuring heart rate variability using commercially available devices in healthy children: A validity and reliability study," *Eur. J. Invest. Health, Psychol. Edu.*, vol. 10, no. 1, pp. 390–404, Jan. 2020.

[77] M. Umair, N. Chalabianloo, C. Sas, and C. Ersoy, "A comparison of wearable heart rate sensors for HRV biofeedback in the wild: An ethnographic study," in *Proc. 25th Annu. Int. CyberPsychol., CyberTherapy Social Netw. Conf.*, 2020.

[78] I. T. Hettiarachchi, S. Hanoun, D. Nahavandi, and S. Nahavandi, "Validation of polar OH1 optical heart rate sensor for moderate and high intensity physical activities," *PLoS ONE*, vol. 14, no. 5, May 2019, Art. no. e0217288.

[79] D. Giles, N. Draper, and W. Neil, "Validity of the polar V800 heart rate monitor to measure RR intervals at rest," *Eur. J. Appl. Physiol.*, vol. 116, no. 3, pp. 563–571, Mar. 2016.

[80] Empatica. *E4 Wristband: Real-Time Physiological Signals: Wearable PPG, EDA, Temperature, Motion Sensors*. Accessed: Nov. 11, 2020. [Online]. Available: https://www.empatica.com/research/e4/

[81] F. Scarpina and S. Tagini, "The stroop color and word test," *Frontiers Psychol.*, vol. 8, p. 557, Apr. 2017.

[82] J. R. Stroop, "Studies of interference in serial verbal reactions," *J. Express Psychol.*, vol. 18, no. 6, p. 643, 1935.

[83] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, "The 'trier social stress test'—A tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, nos. 1–2, pp. 76–81, 1993.

[84] R. Castaldo, L. Montesinos, P. Melillo, C. James, and L. Pecchia, "Ultra-short term HRV features as surrogates of short term HRV: A case study on mental stress detection in real life," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, pp. 1–13, Dec. 2019.

[85] H. J. Baek, C.-H. Cho, J. Cho, and J.-M. Woo, "Reliability of ultra-short-term analysis as a surrogate of standard 5-min analysis of heart rate variability," *Telemedicine e-Health*, vol. 21, no. 5, pp. 404–414, May 2015.

[86] R. Pernice, I. Member, M. Javorka, J. Krohova, B. Czippelova, Z. Turianikova, A. Busacca, and L. Faes, "Comparison of short-term heart rate variability indexes evaluated through electrocardiographic and continuous blood pressure monitoring," *Med. Biol. Eng. Comput.*, vol. 57, no. 6, pp. 1247–1263, Jun. 2019.

[87] M. P. Tarvainen, J.-P. Niskanen, J. A. Lipponen, P. O. Ranta-Aho, and P. A. Karjalainen, "Kubios HRV–heart rate variability analysis software," *Comput. Methods Programs Biomed.*, vol. 113, no. 1, pp. 210–220, 2014.

[88] D. Hernando, S. Roca, J. Sancho, Á. Alesanco, and R. Bailón, "Validation of the apple watch for heart rate variability measurements during relax and mental stress in healthy subjects," *Sensors*, vol. 18, no. 8, p. 2619, Aug. 2018.

[89] D. J. Plews, B. Scott, M. Altini, M. Wood, A. E. Kilding, and P. B. Laursen, "Comparison of heart-rate-variability recording with smartphone photoplethysmography, polar H7 chest strap, and electrocardiography," *Int. J. Sports Physiol. Perform.*, vol. 12, no. 10, pp. 1324–1328, Nov. 2017.

[90] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 3, pp. 230–236, Mar. 1985.
[91] J. A. Lipponen and M. P. Tarvainen, "A robust algorithm for heart rate variability time series artefact correction using novel beat classification," *J. Med. Eng. Technol.*, vol. 43, no. 3, pp. 173–181, Apr. 2019.
[92] S. Ranganathan, K. Nakai, and C. Schonbach, *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*. Amsterdam, The Netherlands: Elsevier, 2018.
[93] H. G. V. Lier, M. E. Pieterse, A. Garde, M. G. Postel, H. A. D. Haan, M. M. R. Vollenbroek-Hutten, J. M. Schraagen, and M. L. Noordzij, "A standardized validity assessment protocol for physiological signals from wearable technology: Methodological underpinnings and an application to the E4 biosensor," *Behav. Res. Methods*, vol. 52, no. 2, pp. 607–629, 2019.
[94] C. S. Bulte, S. W. Keet, C. Boer, and R. A. Bouwman, "Level of agreement between heart rate variability and pulse rate variability in healthy individuals," *Eur. J. Anaesthesiol.*, vol. 28, no. 1, pp. 34–38, Jan. 2011.
[95] E. Schrödl, S. Kampusch, B. D. Razlighi, V. H. Le, J. C. Széles, and E. Kaniusas, "Feasibility of pulse rate variability as feedback in closed-loop percutaneous auricular vagus nerve stimulation," *Vibroengineering PROCEDIA*, vol. 26, pp. 35–39, Sep. 2019.
[96] A. Tonacci, L. Billeci, E. Burrai, F. Sansone, and R. Conte, "Comparative evaluation of the autonomic response to cognitive and sensory stimulations through wearable sensors," *Sensors*, vol. 19, no. 21, p. 4661, Oct. 2019.
[97] D. Giavarina, "Understanding bland altman analysis," *Biochemia Medica*, vol. 25, no. 2, pp. 141–151, 2015.
[98] D. G. Altman and J. M. Bland, "Measurement in medicine: The analysis of method comparison studies," *J. Roy. Stat. Soc. D, Stat.*, vol. 32, no. 3, pp. 307–317, 1983.
[99] J. M. Bland and D. G. Altman, "Measuring agreement in method comparison studies," *Stat. Methods Med. Res.*, vol. 8, no. 2, pp. 135–160, Jun. 1999.
[100] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Int. J. Nursing Stud.*, vol. 47, no. 8, pp. 931–936, Aug. 2010.
[101] J. M. Bland, D. G. Altman, and D. S. Warner, "Agreed statistics: Measurement method comparison," *J. Amer. Soc. Anesthesiol.*, vol. 116, no. 1, pp. 182–185, 2012.
[102] *Heart Rate Meters, and Alarms*, Standard ANSI/AAMI EC13, Cardiac Monitors, 2002.
[103] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet*, vol. 327, no. 8476, pp. 307–310, 1986.



MUHAMMAD UMAIR is currently a Senior Research Associate with the ImaginationLancaster, Lancaster Institute for the Contemporary Arts. Previously, he worked as a Research Associate with the School of Computing and Communication, Lancaster University. His research interests include the field of human–computer interaction and interactive systems. He works on topics around digital health, physical computing, sensors, and the IoT systems. His work combines human-centered design and interaction concepts with new interactive systems, machine learning, toolkits, and interaction techniques. He has published articles at CHI, DIS, and UbiComp, and his Ph.D. research received media coverage with an overall reach of 6.5 million across the world.



NIAZ CHALABIANLOO received the master's degree in computer engineering from Middle East Technical University, Turkey. He is currently pursuing the Ph.D. degree with Boğaziçi University. He is also an Early Stage Researcher for the Marie Sklodowska-Curie H2020 AffecTech Project. His current research interests include data science, pervasive health, and ubiquitous computing.



CORINA SAS received the Ph.D. degree in human–computer interaction from University College Dublin. She is currently 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. Her research is in the area of technologies for well-being and health. She published over 150 papers, and her work received extensive media coverage as well as four Honourable Mention Awards at CHI and DIS. She has been an investigator on grants totaling over £14.7 million and currently coordinates the AffecTech: Personal Technologies for Affective Health, a Marie Sklodowska-Curie Innovative Training Network (2017–2021) funded by the European Commission. She is also part of the Editorial Boards of the *ACM Transactions in Human-Computer Interaction, and Taylor and Francis Human-Computer Interaction Journal*.



CEM ERSOY (Senior Member, IEEE) 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 of computer engineering 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.