Analysis of Psychophysiological Responses Using Heart Rate Variability: Towards Real-Time Affect Recognition

Mariam Ali Salem Bahameish

2022

Submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy

School of Electronics Engineering and Computer Science

Queen Mary University of London

Statement of Originality

I, Mariam Ali Salem Bahameish, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged and my contribution indicated. Previously published material is also acknowledged. I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material. I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis. I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Mariam Bahameish

6th September, 2022

Collaboration: As part of my sponsorship affiliation with the Qatar National Research Fund (a member of Qatar Foundation), the heart rate variability biofeedback study was conducted in collaboration with Dr Dena Althani at Hamad Bin Khalifa University (HBKU; itself a member of Qatar Foundation) in Qatar during the COVID-19 period. The principal investigator of the study, Dr Dena Althani, assisted in data collection and reviewed the research study before submitting the ethical application to the Institutional Review Board at HBKU. My contributions included study conception and design, data collection, data analysis, and interpretation of results.

Abstract

Recent technological developments have provided innovative means for promoting health and well-being through physiological response monitoring. Heart rate variability (HRV) has arisen as a promising physiological indicator of mental health. This research contributes towards these efforts by investigating the short-term effects of increased HRV using a biofeedback exercise (paced breathing) on affective states and physiological measures to facilitate the development of real-time affect recognition systems.

To enable the analysis of high-quality HRV data in real-time applications, the first study examined the reliability of automatic filtering techniques using an open-source implementation. The outcomes of this study provided a flexible control for HRV signal filtering parameters and served as the basis for the analyses in the following studies. Subsequently, the second study investigated the minimum reliable window for HRV signals based on the conditions under which the data were collected. The findings suggested that HRV measures can be analysed in segments of less than 5 minutes in all conditions. Additionally, the minimum segment differed in paced breathing compared to resting and stress. Given the physiological influence of paced breathing, the third study examined the short-term effects of a heart rate variability biofeedback (HRVB) intervention on a range of affective states (e.g., relaxation, stress), working memory, and physiological data. The findings showed a significant improvement in working memory and relaxation levels following the intervention. The last study leveraged the major findings of the previous two studies to develop robust predictive models that identified stress using supervised learning algorithms.

Overall, this research demonstrates that a single HRV biofeedback session mediates physiological responses and that this mediation can be measured across a range of affective states. Moreover, it shows that stress levels can be robustly recognised using supervised learning algorithms. This research also lays the groundwork for the potential employment of HRV in real-time applications to predict affective states.

بسم الله الرّحمٰن الرّحيم

(وما توفيقي إلّا بالله عليه توكّلت وإليه أنيب)

[هود: ٨٨]

In the name of Allah, most benevolent, ever-merciful.

And my success is not but through Allah.
Upon Him I have relied, and to Him I return.

[Hud: 88]

Acknowledgements

All praise is due to Allah for granting me the strength, passion and perseverance to complete this research. Allah the Almighty has sustained me in a way for which I will be eternally grateful. He has bestowed upon me countless blessings and surrounded me with incredibly supportive people.

I am deeply grateful to my supervisor, Dr Tony Stockman, for his invaluable guidance, support and encouragement throughout this journey. His expertise and mentorship have greatly contributed to my academic and personal development. I would also like to thank the members of my committee, Dr Jesus Requena and Dr Akram Alomainy, for their valuable feedback and critiques of the various studies comprising this research. I extend my gratitude to my examiners, Dr Stefan Poslad and Dr Stuart Cunningham, for their constructive comments and insightful suggestions during the oral examination.

I offer special thanks to Dr Dena Althani for her assistance in facilitating the experimental study at Hamad Bin Khalifa University in Qatar. I am also indebted to Dr Osama Halabi, who sparked my interest in research and introduced me to the field of human-computer interaction. I am also grateful to my colleagues and friends in our research group, particularly Nida Aziz and Sebastian Ruiz, for their helpful discussions and ideas.

Words alone cannot express my gratitude to my family for their love, motivation and unwavering support throughout the years. My parents, my sisters and my brother have been a constant source of inspiration and encouragement. I am also thankful to my family relatives and friends in Qatar for their support and motivation despite the distances. I am especially indebted to my grandmother for her love and constant prayers. Many thanks to all of my friends in London, especially Maryam Al-Abdulla, Maha Al-Barrak and Hayat Alfagham, for their wonderful friendship and kindness and for the unforgettable memories we have shared together.

Finally, I would like to express my sincere appreciation to the Qatar National Research Fund for sponsoring and funding my postgraduate studies. I am also grateful for the assistance provided by their technical team and the Qatar Research Leadership Program's administrative team. Special recognition to Dr Ayman Bassil and Maria Susi for their continuous support and guidance.

You can't connect the dots looking forward;
you can only connect them looking backwards.

So you have to trust
that the dots will somehow
connect in your future.

— STEVE JOBS

Contents

List of Figures

List of Tables

Acronyms

ANS	autonomic nervous system
CI	confidence interval
CR	correct responses
CV	cross validation
DASS-21	Depression Anxiety Stress Scale
DBP	diastolic blood pressure
DT	decision trees
ECG	electrocardiogram
FFT	fast Fourier Transform
GAD-7	Generalized Anxiety Disorder Scale
HCI	human-computer interaction
HRV	heart rate variability
HRVB	heart rate variability biofeedback
ICC	intraclass correlation coefficient
IPAQ	International Physical Activity Questionnaire
KNN	k-nearest neighbours
LOGO	leave-one-group-out
LR	logistic regression
MAE	mean absolute error
ML	machine learning
MLL	multilevel linear
NB	Naive Bayes
PANAS	Positive and Negative Affect Schedule
PPG	photoplethysmography

PSNS	parasympathetic nervous system
PSQI	Pittsburgh Sleep Quality Index
RCT	randomised controlled trial
RF	resonant frequency
RFC	random forest classifier
RSA	respiratory sinus arrhythmia
RT	reaction time
SBP	systolic blood pressure
SNS	sympathetic nervous system
SVM	support vector machine
TSST	Trier Social Stress Task
UST	ultra-short-term

CHAPTER 1

Introduction

Mental health extends beyond the absence of mental illness or psychological disorder; it is a fundamental and integral component of overall health (World Health Organization [WHO], 2001). According to the WHO Constitution, mental health can be described as follows:

> A state of well-being in which the individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to his or her community (WHO, 2004, p. 10).

There is now a general consensus that positive mental health and well-being can influence a wide range of behavioural, emotional, and psychological components, including cognitive ability, emotional regulation, physical health and fitness, productivity, quality of life, and sleep quality (Friedli, 2009). These components affect the stability—homeostasis—of the autonomic nervous system (ANS), which comprises two branches: sympathetic (fight-or-flight) and parasympathetic (rest-and-digest; Kemp & Quintana, 2013).

In 1932, a novel direction was established by Cannon (1932, as cited in Kemeny, 2003), indicating that perturbations in the ANS are linked to various psychological and emotional responses. This phenomenon stimulates sympathetic and parasympathetic activities, which can then be captured through changes in physiological measures (e.g., an increase in blood pressure [BP] or heart rate [HR]). More recently, the relationship between disruptions in the ANS and physiological signs of psychological and behavioural disorders has been delineated by investigations in the domains of psychiatry (Alvares et al., 2016), depression (Chalmers et al., 2014), emotions (Kreibig, 2010; Levenson, 1992), and stress (Kemeny, 2003; Won & Kim, 2016).

This research work provides a novel perspective that contributes towards the improvement of mental well-being through the use of sensors to examine changes in cardiac activity and explore robust machine learning (ML) techniques to identify stress levels. The following sections present the motivation behind this research, primary research questions, summary of the research contributions, overview of the thesis structure, and associated publications.

1.1 Motivation

Affective Computing

Over the last few decades, the construct of mental health has evolved, and these changes have been reflected in the literature. Recent technological breakthroughs have provided innovative means to promote mental health and well-being (Curtiss et al., 2021; Greene et al., 2016): namely, the monitoring and management of mental states using psychophysiological measures to document a meaningful relationship between psychological and physiological responses (Edgar et al., 2009).

A seminal work in this field is *Affective Computing*, published in 1997 and written by Rosalind Picard of the MIT Media Lab (Picard, 1997). This text created tremendous opportunities for developing effective solutions to recognise, process, and interpret human feelings using the intersection of human-computer interaction (HCI) and ML research. Today, affective computing is a broad field that focuses on the recognition of affective states (e.g., cognition, emotion, relaxation, stress) using computers and wearable sensors (McDuff et al., 2016; Papadopoulou et al., 2019; Umematsu et al., 2020).

Pioneering work by Healey and Picard (2005) showed the feasibility of recognising drivers' stress levels in a real-world automotive environment using physiological data obtained from cardiac activity, muscle activity, respiration, and skin conductance. A plethora of psychophysiological studies emerged focusing on the identification of affective states using skin conductivity exclusively (Greco et al., 2016; Kurniawan et al., 2013; Liapis et al., 2015). However, skin conductance data only capture the stimulation of sympathetic activity provoked by stressful events (Anusha et al., 2018). Stimulation of parasympathetic activity is essential to bring the body into a relaxed state. Thus, the ability to improve, rather than merely identify, affective states through the stimulation of parasympathetic activity remains an active area of research.

Cardiac Activity

Recent studies have incorporated HR and heart rate variability (HRV) data because the cardiovascular system manifests parasympathetic information in addition to sympathetic responses (Kim et al., 2018). HRV is determined by the time interval between two consecutive heartbeats, which provides insights into ANS imbalance (Berntson et al., 1997). Under normal conditions, parasympathetic and sympathetic responses are in a balanced state. When sympathetic activity is stimulated, HR increases and the variation between heartbeats decreases, resulting in reduced HRV (Wehrwein et al., 2016). A lower HRV value indicates that a person is experiencing stress, whether mental or physical. On the other hand, the stimulation of parasympathetic activity decreases HR, resulting in higher HRV levels (Ernst, 2017).

An association between higher resting HRV measures and better emotion recognition was demonstrated by Quintana et al. (2012), indicating a relationship between the ANS and cognitive processes. While exploring ways to improve HRV, biofeedback through paced breathing exercises emerged as a promising approach for enabling an individual to increase HRV by activating the parasympathetic response. Further, heart rate variability biofeedback (HRVB) is an effective technique for building resilience and improving mental health and well-being in the long term (Gevirtz, 2013; Lehrer & Gevirtz, 2014). In a recent meta-analytic study, Lehrer et al. (2020) reported a significant and small-to-medium effect size regarding the efficacy of multiple-session HRVB in improving a wide variety of physical (e.g., asthma) and psychological symptoms (e.g., anxiety, depression).

More recent attention has focused on the short-term effects of a single HRVB session, and the preliminary results demonstrated significant benefits in the context of momentary stress management (Prinsloo et al., 2013; Yu et al., 2018b), inhibitory control (Laborde et al., 2019b), mood improvement (Steffen et al., 2017), and food-craving control (Meule & Kübler, 2017). Overall, these studies highlight the positive aspects of a single HRVB session; however, further research is needed to examine the short-term effects of HRVB practice on psychophysiological responses.

Real-Time Applications

The provision of real-time applications for continuously monitoring physiological data and providing immediate feedback to the user has a considerable impact on improving critical mental and physical health decisions (Abdullah & Choudhury, 2018). For instance, leveraging the cardiac sensing abilities of wearable devices can provide indicative early insights regarding the symptoms of anxiety and depression disorders (Chalmers et al., 2014), bipolar disorder and schizophrenia (Quintana et al., 2016a), and post-traumatic stress disorder (Tan et al., 2011).

Within the context of HRV, the analysis is often performed on long-term (24 h) or short-term (5 min) recordings (Malik et al., 1996). In the last few decades, there has been a surge of interest in conveying momentary information about the parasympathetic and sympathetic activities in clinical environments (Kasaoka et al., 2010) and research settings (Castaldo et al., 2019; Shaffer et al., 2016). Therefore, ultra-short-term (UST) analysis has emerged as an effective approach to segment HRV recordings in windows of less than 5-min using a series of overlapping segments (Pecchia et al., 2018; Shaffer et al., 2020). These segments are updated in real-time via small time increments to provide information about the dynamic and momentary changes of HRV (Zhang et al., 2015).

It could be argued that UST analysis is in fact not a real-time approach because a short delay may occur due to the analysis's reliance on ultra-short segments (He et al., 2019). However, it is a generally acknowledged term because the ultimate aim is to offer instantaneous and momentary information based on the real-time acquisition of the HRV data from clinical instruments or wearable devices (Jiang et al., 2017; Kasaoka et al., 2010; Pecchia et al., 2018; Shiraishi et al., 2018). Therefore, the minimum reliable UST segment should be determined to obtain a close approximation of real-time analysis based on the condition under which the HRV data have been collected.

1.2 Research Questions

The overarching aim of this research is to investigate the short-term effects of increased HRV using biofeedback on affective states and physiological measures for eventual deployment in real-time recognition systems. To achieve this objective, the following research question was posited:

> **RQ:** How does a single HRVB session using paced breathing mediate physiological responses across a range of affective states, and can these affects be robustly recognised by supervised learning algorithms?

Accordingly, there are four main areas of focus and each area is addressed by an associated research sub-question (SRQ), as follows:

1. Effective research in the psychophysiological domain begins by ensuring that the relevant signals have sufficient quality and reliability. In the context of HRV analysis for real-time applications, these concerns were addressed by the following sub-question and are discussed at length in Chapter 4:

   > **SRQ1:** What signal preprocessing algorithms are necessary for a reliable real-time HRV analysis?

2. For moment-by-moment analysis, UST segments have been used to analyse HRV signals in periods of less than 5 min to ease deployment in real-time applications. Thus, the necessary requirements for UST analysis under resting and non-resting conditions were examined by the following sub-question and are addressed in Chapter 5:

   > **SRQ2:** What are the requirements for a reliable real-time HRV analysis using UST segments under resting, stress, and paced breathing conditions?

3. To determine whether a single-paced breathing HRVB session would have a positive impact on affective states and physiological measures, the following sub-question was posed and is further discussed in Chapter 6:

   SRQ3: What is the effect of a single paced breathing session on affective states (cognition, relaxation, stress) and physiological responses (HRV and BP)?

4. Finally, it was necessary to determine whether the changes in affective state examined in Chapters 5 and 6 could be correctly classified by robust supervised learning algorithms, thereby facilitating deployment in real-time recognition applications. These considerations were expressed in the following sub-question and are explicated in Chapter 7:

   SRQ4: How can robust supervised learning algorithms recognise stress and relaxed states for eventual deployment in real-time systems?

1.3 Contributions

The novelty of this research is presented in relation to the major contributions that it has produced. Further details about the contributions are discussed in Chapter 8.

1. **Preprocessing of Heart Rate Variability Data**
   To enable batch processing and the deployment of high-quality HRV data in real-time applications, automatic filtering techniques were implemented to detect and correct existing artefacts using a flexible open-source environment. Accordingly, a real-time framework was developed for integrating the automatic filtering approach with online HRV data acquisition.

2. **Ultra-Short-Term Analysis of Heart Rate Variability Data**
   To simulate real-time HRV data acquisition, a concurrent validity assessment of a standard 5-min HRV signal was used to establish the minimum reliable segment of HRV analysis based on the conditions under which the data were acquired. Further, the influence of stress and paced breathing was examined, with special consideration given to measurement consistency across UST segments.

3. **Impact of Heart Rate Variability Biofeedback on Affective States**
   Following HRV preprocessing and UST analysis, the influence of HRVB on psychophysiological responses was investigated. The results demonstrate the short-term effects of a single HRVB paced breathing session on a range of affective states (e.g., attentiveness, fatigue, mood, serenity, stress), executive function (i.e., a working memory task), and physiological data (i.e., HRV and BP).

4. **Robust Techniques for Stress Recognition**
   Robust strategies for ML affect recognition based on physiological measures were proposed to address current limitations in the literature. With an emphasis on stress as a prominent affective state, these robust techniques were used to predict stress levels from HRV data using supervised learning algorithms.

1.4 Thesis Outline

The remainder of this thesis is structured as follows (see Figure 1.1):

- **Chapter 2** provides background information and discusses relevant work in the areas of affective states, HRV, and biofeedback, thereby delineating the foundational knowledge for this research.

- **Chapter** 3 elaborates on the methods employed in this research by describing the experimental design, data collection, data analysis, and ethical considerations.

- **Chapter** 4 describes a fundamental study for filtering HRV data via the detection and correction of existing artefacts in an open-source flexible environment to facilitate batch processing and real-time analysis.

- **Chapter** 5 presents an exploratory study on the minimum reliable UST segment for HRV analysis under resting, stress, and paced breathing conditions.

- **Chapter** 6 presents an investigation of the short-term effects of HRVB via paced breathing on affective states, executive function, and physiological responses.

- **Chapter** 7 leverages the datasets and major findings from the two preceding chapters to present robust stress recognition models using supervised ML algorithms.

- **Chapter** 8 concludes by summarising the key findings, discussing relevant contributions in light of the research questions, outlining study limitations, and proposing potential directions for future research.

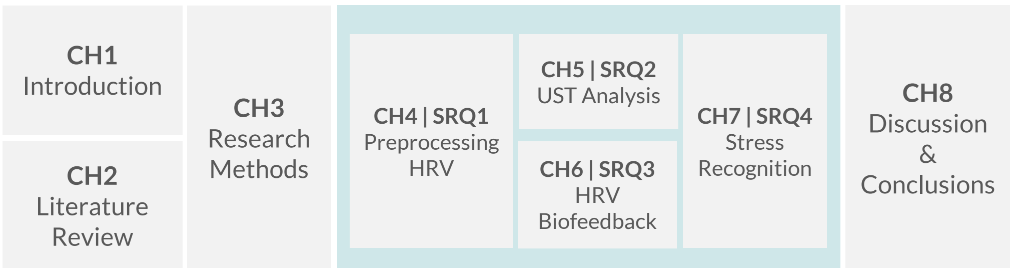


Figure 1.1: Thesis Outline

1.5 Publications and Presentations

Parts of this thesis were published or presented as posters in the following contexts:

- Bahameish, M., & Stockman, T. (2020). Fundamental Considerations of HRV Analysis in the Development of Real-Time Biofeedback Systems. *2020 Computing in Cardiology Conference (CinC)*, *47*. https://doi.org/10.22489/CinC.2020.078
- Bahameish, M., & Stockman, T. (2019a). The analysis of heart rate variability measures in ultra-short-term window segments. *EECS Research Open Day, Queen Mary University of London, UK*
- Bahameish, M., & Stockman, T. (2019b). Facilitating the control of stress levels in real-time as manifested in measures of heart rate variability. *The 1st International Conference on Visualization and Computer-Human Interaction (VisCHI), Doha, Qatar*

CHAPTER 2

Literature Review

This chapter reviews the literature related to the three primary domains explored in this thesis: affective states, heart rate variability (HRV), and heart rate variability biofeedback (HRVB). It begins with a description of the affective states examined as well as their relationship to physiological responses, with the ultimate goal of utilising these responses in machine learning-based recognition applications. Moreover, it delineates the physiological and theoretical background of HRV, including methods for signal preprocessing and analysis, followed by an explanation of HRVB practice as an approach for improving mental well-being.

2.1 Affective States

Recent trends in the field of affective computing have led to a proliferation of studies aimed at improving health and well-being (Curtiss et al., 2021; Hasnul et al., 2021; Schmidt et al., 2019). Although the term "affective" was exclusively associated with emotions when the field was first established (Picard, 1997), it is now regarded as an umbrella term that covers a wide range of states, including cognition, emotion, and stress (Picard, 2016).

The following sections describe the affective states examined in this research (i.e., stress and executive function) as well as their relationship to physiological responses, followed by a description of wearable devices that can detect such responses. Lastly, previous research in the domain of affect recognition using ML is discussed.

2.1.1 Stress

Feeling stressed or being subjected to a stressful event is an inevitable experience of life. Although there is no global consensus on the definition of stress, a widely accepted definition among researchers was coined by Selye (1978): "a nonspecific response of the body to any demand, whether it is caused by, or results in, pleasant or unpleasant conditions" (p. 74). Pleasant or unpleasant conditions result in two distinct forms of stress: positive "eustress" and negative "distress". Eustress is accompanied by a feeling of extreme joy or excitement; in contrast, distress is related to physical or emotional suffering. Selye (1976, as cited in Fink, 2017) proffered an additional definition of stress for behavioural science contexts: "perception of threat, with resulting anxiety, discomfort, emotional tension, and difficulty in adjustment" (p. 4). These definitions are closely interwoven as stress causes the human body to shift from a calm to aroused state (Healey & Picard, 2005). For the sake of clarity, the term *stress* in this thesis hereafter refers to acute psychological distress elicited in a laboratory setting.

According to Cannon (1932, as cited in Fink, 2017), internal bodily reactions to stress reflect activation of the fight-or-flight response, which disrupts the body's homeostasis (i.e., the stabilised state of the ANS; see Section 2.2.2). Hence, the disruption of homeostasis due to a psychologically demanding task causes mental arousal, which is interpreted as a stress response and can be observed in physiological changes (Campbell & Ehlert, 2012; Crosswell & Lockwood, 2020; Dimsdale, 2008; Kemeny, 2003; Skoluda et al., 2015; Steptoe & Vogele, 1991). Stress can be further categorised according to its duration. Acute stress is a short-term feeling of pressure related to anticipated distressing events, and it evokes such physiological responses as accelerated HR, muscle tension, rapid breathing, and sweat gland activation. Chronic stress, on the other hand, is long-term stress resulting from exposure to high-pressure situations (e.g., demanding occupations) and has an adverse effect on mental and physical health (Greene et al., 2016).

Psychological stress can be investigated in both laboratory settings (Steptoe & Vogele, 1991) and real-life environments, such as academic examinations (Melillo et al., 2011) or firefighting scenarios (Smith et al., 2001). Laboratory settings offer a controlled experimental environment where stress responses can be safely provoked for the purposes of inferring statistically significant relationships among the investigated variables. Early research established reliable stress protocols for such physiological reactions as hormonal and cardiac reactivity, including the Trier Social Stress Test (TSST) protocol (Kirschbaum et al., 1993).

However, it should be noted that lab-based experiments are susceptible to limitations in ecological validity (Hayes, 2021). Further, there is a distinct lack of empirical research investigating approaches to improve, rather than merely identify, stress levels. To address this gap, the present research sought to induce and alleviate stress using psychological tests and paced breathing practices, respectively.

2.1.2 Executive Function

Six major domains of cognitive function have been postulated by the American Psychiatric Association: namely, complex attention, executive function, language, learning and memory, perceptual-motor function, and social cognition (Sachdev et al., 2014). Each domain covers a variety of cognitive processes, as depicted in Figure 2.1.

Studies on executive function represent a growing field due to its direct association with activities of daily living (Chan et al., 2008; Starcke et al., 2016). Executive function refers to the higher cognitive skills underpinning self-control

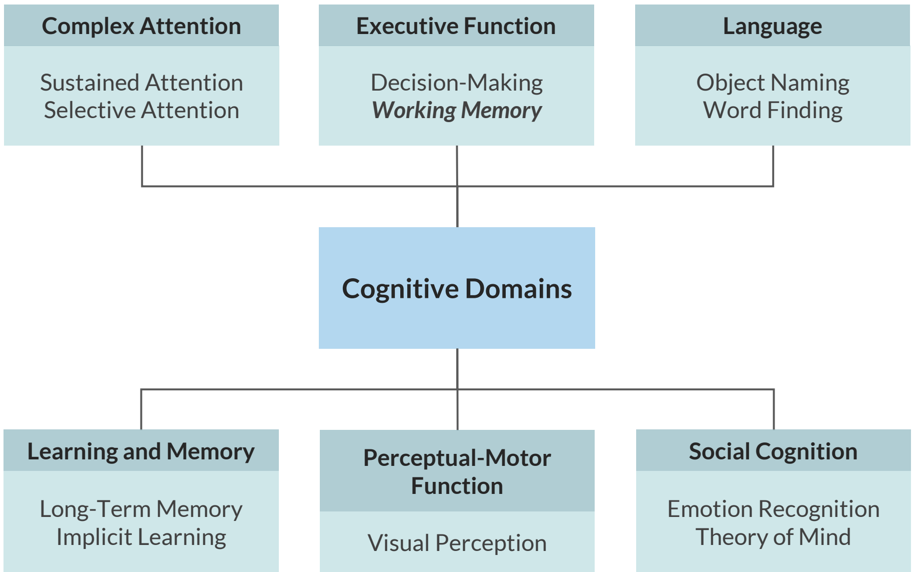


Figure 2.1: Cognitive Domains
Note. Originally published in Sachdev et al. (2014)

and goal-directed behaviour, including decision-making, problem-solving, and self-regulation. It comprises three primary areas: inhibitory control, working memory, and cognitive flexibility (Diamond, 2013). Inhibitory control involves the capacity to self-regulate one's attention, actions, and emotions. Working memory involves the retention of information for a limited amount of time as it is being mentally processed. Cognitive flexibility refers to the mental ability to adapt to new situations or changes, and it is based on inhibitory control and working memory.

From a neuroscientific perspective, the executive function processes are predominantly located in the prefrontal cortex of the frontal lobe and supported by connected brain structures, such as the amygdala (Blair, 2017). Benarroch (1993) established that central autonomic control is a bidirectional interaction between the central nervous system (CNS), which consists of the brain, and the ANS, which governs the organs of the body. Subsequently, Thayer and Lane (2000) posited the neurovisceral integration model, which emphasises the connection between the prefrontal cortex and the cardiovascular system via the vagus nerve of the ANS (see Section 2.2.2). This model effectively disentangled the relationship between the brain, heart, and cognitive processes. Further details of the neurovisceral integration model are discussed in Chapter 6 (see Section 6.2).

Executive dysfunction poses considerable challenges to the capacities for concentration, planning, self-control, and task completion. These challenges can also manifest as deficits in other cognitive domains, such as learning and memory and social cognition (Groden et al., 2005). Hence, various psychological tests have been developed to assess for deficits in executive function, such as the Stroop Color and Word Test (SCWT) to assess inhibitory control, backward digit span (BDS) and N-back tasks to assess working memory, and task switching to assess cognitive flexibility. The details regarding the protocol for each task are described in Diamond (2013) and, more recently, Friedman and Robbins (2022). As this research sought to assess the capacity of working memory, the N-back task was selected to evaluate cognitive performance (see Chapter 6); further, a variation of the BDS was employed as an arithmetic component of the stress inducer in the TSST (see Chapter 5).

2.1.3 Wearable Sensors

Researchers have long striven to develop ubiquitous wearable technologies for the purposes of health monitoring and improvement (Dunn et al., 2010; Lisetti et al., 2003; Luneski et al., 2010; Picard, 2009). An eminent advantage of wearable sensors is that they provide non-invasive direct physical contact with the user for long-term periods (Picard & Healey, 1997). The data collected can yield insights regarding the user's activity patterns, thereby enhancing future deployment for physical and mental health management.

Further, the physiological signals for health monitoring (e.g., brain activity, cardiac activity, muscle activity, respiration, skin conductance) can be collected via embedded sensors in wearable devices (Greene et al., 2016). The integration of wearable sensors with emotion and stress detection has attracted considerable attention given that affective states primarily influence the ANS, as reflected by the physiological data (Kreibig, 2010). Consumer-based wearable devices can collect valuable information, such as the Garmin watch which provides a stress score on a scale of 0-100 based on cardiac activity (Garmin, 2022). However, both the accuracy and reliability of such measurements are questionable as it is impossible to access the raw data (Hinde et al., 2021). These limitations preclude the generation of in-depth insights about the intrinsic physiological data gathered as well as the identification of other affective states. The following section provides a summary of the relevant research undertaken in the field of affective computing to develop emotion and stress recognition systems.

2.1.4 Recognition Systems

Affect recognition plays a pivotal role in discerning the internal bodily feelings (e.g., fear, happiness, stress) that influence mental health and well-being (Picard et al., 2001). Traditionally, mental health symptoms have been assessed using clinically validated self-reported questionnaires, such as the Patient Health Questionnaire (PHQ-9) for depression assessment (Kroenke et al., 2001). However, these questionnaires are susceptible to subjective bias as respondents may provide inaccurate or imprecise answers (Demetriou et al., 2015). Fortunately, questionnaires can be supported by physiological data to provide a reliable approach for determining an individual's mental state.

The concept of inferring mental states from physiological data is not new. It dates back to the 1920s, when the lie detector was invented by sensing changes in BP, breathing, and HR (Synnott et al., 2015). However, the recent progress in wearable sensor technology has facilitated the development of more advanced affect recognition and health monitoring systems. Indeed, continuous monitoring of physiological data has the potential to identify early warning signs for mental disorders (Tutunji et al., 2021). Picard envisaged that continuous monitoring via wearable sensors would alleviate the risk for mental illness by detecting the phase prior to the transition from "healthy" to "unhealthy", "normal" to "depressed", and "resilient" to "vulnerable" (Mertz, 2016).

Given the complexity of psychophysiological responses, myriad studies have examined the development of affect detection and recognition prototypes using ML, including supervised, unsupervised, and deep learning. The ML discipline is concerned with the development of statistical models that use existing data to "learn" underlying patterns and behaviours in order to predict future outcomes (Mitchell, 1997). ML is, thus, a promising approach for classification and recognition that has achieved remarkable success in a wide variety of fields, particularly clinical applications (Davenport & Kalakota, 2019; Yan et al., 2019).

At the beginning of the century, Picard et al. (2001) performed pioneering research that shifted the focus away from affect recognition using facial and verbal expressions and towards affect recognition using physiological responses. The data collected from a single participant over several weeks yielded results with a classification performance of 81% for eight elicited emotions derived from breathing, cardiac activity, muscular activity, and skin conductance. This research paved the way for later studies utilising ML algorithms with data obtained from multiple participants to recognise affective states, including emotion recognition (Egger et al., 2019; Kim et al., 2004), fear detection (Bălan et al., 2019; Ihmig et al., 2020), and stress classification (Healey & Picard, 2005; Zhai & Barreto, 2006). Although attempts have been made to apply

ML algorithms within the field of affective computing, limited datasets have impeded the processes of learning and prediction. Chapter 7 goes into further detail regarding these limitations as well as potential mitigation strategies.

In a comprehensive review of affect recognition, Schmidt et al. (2019) examined the detection of several affective states, including emotion, excitement, frustration, happiness, relaxation, and stress. Most of the studies (34 out of 46) focused on identifying stress levels (16 studies) and emotional states (18 studies), with the two-dimensional circumplex model (i.e., valence-arousal) used for the latter (Lang et al., 1997; Russell, 1979). The results highlight the use of various physiological signals in the reviewed studies: 40 used cardiac activity, 35 used skin conductivity, 15 used miscellaneous signals (e.g., accelerometer data, muscle activity, respiration, temperature), and seven used brain activity. Skin conductance was measured using electrodermal activity, sometimes known as the galvanic skin response, which quantifies the fluctuation of electrical activity by applying a low voltage to the skin. Electrical activity often increases as the skin receives innervating impulses from the brain, as evidenced by sweat secretion (Boucsein, 2012). Because skin conductivity is entirely governed by the fight-or-flight response (i.e., the sympathetic branch of the ANS), it is primarily used to identify arousal in psychological or physiological responses (Kim & André, 2008).

On the other hand, HRV captures the activity of both branches of the ANS with a predominance of the parasympathetic activity. Thus, the data gathered from participants under resting conditions indicate the stimulation of parasympathetic activity (i.e., a state of calmness and relaxation) and can provide insights regarding positive mental well-being. In contrast, reduced parasympathetic activity indicates a threat to homeostasis (e.g., an exposure to stressful events). The following section presents the physiological background of HRV and provides an overview of the methods used for signal preprocessing and analysis.

2.2 Heart Rate Variability

HRV is a measure of the time variation between successive heartbeats in milliseconds (Berntson et al., 1997), which is commonly referred to as the RR interval or interbeat interval (IBI). HRV is widely used as a non-invasive indicator of the balance in the ANS (Sztajzel, 2004), which reflects the status of mental well-being and physical health.

2.2.1 Physiological Background

A heartbeat, or cardiac cycle, is initiated when the sinoatrial (SA) node generates an electrical impulse that spreads through the upper heart chambers (atria), causing them to eject blood into the lower heart chambers (ventricles). The ventricles then contract to pump blood to the body's organs. In the process, electrical activity passes through the atrioventricular (AV) node, which controls blood flow by slowing down electrical impulses before they can reach the ventricles (Kitney & Rompelman, 1980). The SA node produces 60-100 electrical impulses per minute and is, thus, considered a natural pacemaker or "the driving force of heart's rhythm" (McDonald, 1980, p. 1). There are two processes that result from the contraction of the atria and ventricles: depolarisation (activation) and repolarisation (recovery).

The electrocardiogram (ECG) is a time-voltage visual representation of the cardiac cycle's events. It can be acquired by attaching two or more electrodes directly onto the skin. There are three main components of the ECG signal: the P wave representing atrial depolarisation, the QRS complex representing ventricular depolarisation, and T wave representing ventricular repolarisation.

HRV is generally calculated from an ECG as a measure of the time between the R-peaks of consecutive QRS complexes: hence the name RR interval.

However, photoplethysmography (PPG) can be used as an alternative method for calculating HRV. It is an optical measurement of light propagation through the skin to determine the rate of blood flow, as controlled by the pumping action of the heart. Similar to the ECG, cardiac cycles appear as peaks in the PPG signal, and the peak-to-peak interval is used to calculate IBI (Allen, 2007). The PPG approach is commonly used in wearable devices because of its simplicity. Nonetheless, it poses reliability concerns in non-stationary conditions (e.g., during physical exercise; Castaneda et al., 2018). Figure 2.2 presents a labelled simulation of synthetic ECG and PPG recordings. Throughout this thesis, RR interval is used to refer to the HRV signal collected from either ECG or PPG.

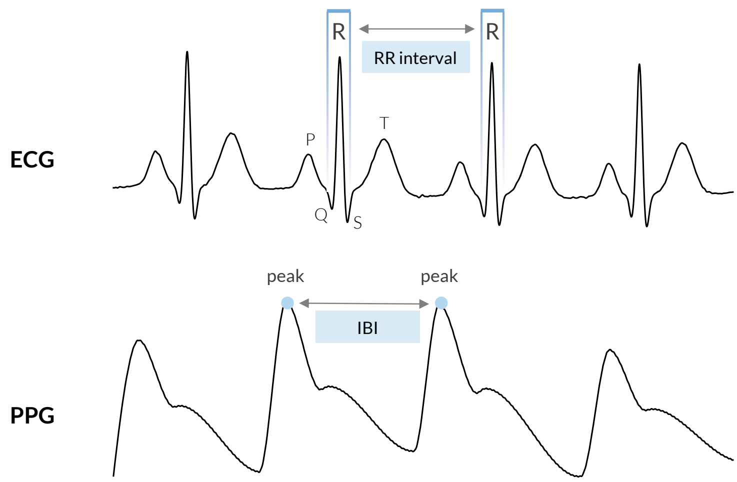


Figure 2.2: A snapshot of synthetic ECG and PPG recordings showing the time variation between successive heartbeats in milliseconds of RR and IBI, respectively

Note. Created with NeuroKit2 Python package (Makowski et al., 2021)

2.2.2 Autonomic Nervous System

The nervous system has two components: the CNS and the peripheral nervous system (PNS). The CNS is composed of the brain and the spinal cord, while the PNS is composed of the nerves that connect the CNS to the body organs (e.g., glands, heart, lungs). A subdivision of the PNS is the ANS, which regulates internal bodily processes by releasing specific enzymes and hormones to control them. Examples of these processes include BP, digestion, HR, and respiration (Kaltsas & Chrousos, 2007).

The ANS bifurcates into sympathetic and parasympathetic branches. The sympathetic nervous system (SNS) stimulates the fight-or-flight response, which is a physiological reaction triggered by stressful or threatening situations. Conversely, the parasympathetic nervous system (PSNS) stimulates the rest-and-digest response, which is associated with relaxed and resting states. Both branches seek to maintain a state of equilibrium known as homeostasis, which is a complex and dynamic stable condition (Bankenahally & Krovvidi, 2016). Within the context of the cardiovascular system, the ANS branches have an antagonistic effect. When the SNS is activated, HR increases; conversely, when the PSNS is activated, HR decreases (Wehrwein et al., 2016). At a resting condition, parasympathetic activity is dominant, resulting in an average HR of 75 beats/min for healthy adults (Shaffer & Ginsberg, 2017).

The vagus nerve is the longest cranial nerve of the ANS, and it carries sensory and motor information between the body organs and brain (Porges, 1995). It also provides parasympathetic innervation of the heart, which causes the HR to slow down. The contribution of vagus nerve activity to the heart is referred to as *vagal tone*, and it can be quantified using HRV. In particular, vagal tone is indexed using respiratory sinus arrhythmia (RSA), which is a natural increase or decrease in HR in response to respiration. Additionally, vagal tone is affected by the baroreflex, which provides rapid feedback regarding BP changes by decelerating the HR in response to an increase in BP, and vice versa, to maintain homeostasis. Modulating vagal tone contributes to the dynamic balance of the ANS, which is vital to both cardiovascular and overall health (Berntson et al., 2009).

2.2.3 Signal Preprocessing

Artefacts (e.g., abnormalities, outliers) in ECG and PPG recordings can have a significant impact on the reliability of HRV analysis (Berntson et al., 1997; Choi & Shin, 2018). The primary sources of artefacts are physiological responses, such as ectopic beats, and technical issues, such as sensor motion. At the ECG level, the artefact is represented as an extra or missed beat (see Figure 2.3; Berntson et al., 1990). These beats are reflected in the HRV signal by a short or long period between successive intervals relative to the normal RR interval for extra and missed beats, respectively.

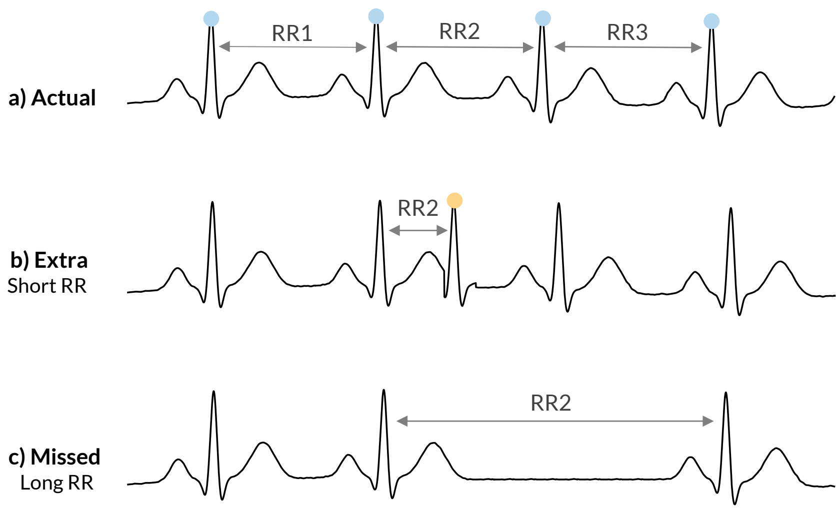


Figure 2.3: A schematic diagram of ECG signals a) without artefacts, b) with an extra beat, and c) with a missed beat.
Note. Created with NeuroKit2 Python package (Makowski et al., 2021)

Two major processes are considered to filter out artefacts for better HRV signal quality: artefact detection and artefact correction. Several techniques for artefact detection have been developed by identifying unexpected intervals based on a fixed arbitrary threshold, such as a 20% difference between RR intervals (Malik et al., 1989) or a 32.5% increase and 24.5% decrease between them (Kamath & Fallen, 1995, as cited in Choi & Shin, 2018; Pichot et al., 2016). Recent fixed detection techniques have employed a modified flexible approach by calculating a local median or mean threshold value for the HRV signal (Tarvainen et al., 2016). Advanced techniques have further focused on adaptive, rather than fixed, approaches by calculating time-varying thresholds (Lipponen & Tarvainen, 2019). Once the artefacts are identified, they can be corrected using deletion, averaging, or interpolation techniques (Peltola, 2012). This research examines HRV preprocessing techniques in detail by assessing the reliability of artefact detection and correction methods for eventual deployment in real-time applications (see Chapter 4).

2.2.4 Heart Rate Variability Analysis

Three standardised analytical approaches for quantifying HRV have been articulated by the Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (Task Force; Malik et al., 1996): time-domain, frequency-domain, and non-linear methods (see Table 2.1; Shaffer & Ginsberg, 2017).

Time-Domain primarily consists of statistical analysis of the HRV signal obtained from ECG or PPG recordings and calculated using indices of mean, standard deviation (SDNN), the absolute differences between successive beats (NNx), the percentage of differences between successive beats (pNNx), and the root mean square of successive differences (RMSSD).

Table 2.1: Heart Rate Variability Measures

Measure	Unit	Mechanism	Description
Time Domain			
Statistical			
MeanRR	ms		Average of NN[1] intervals
SDNN	ms	SNS[2] and PSNS[3]	Standard deviation of NN intervals
RMSSD	ms	PSNS	Root mean square of successive differences between adjacent NN intervals
NN50	–	PSNS	Number of differences that differ by 50 ms between adjacent NN intervals
pNN50	%	PSNS	Percentage of differences that differ by 50 ms between adjacent NN intervals
Geometric			
TRI	–		Integral of the density of the RR interval histogram divided by its height
TINN	ms		Baseline width of the RR interval histogram
Frequency Domain			
ULF power	ms^2	SNS and PSNS[*]	Power of the ultra-low frequency band: (< .003 Hz)
VLF power	ms^2	SNS and PSNS	Power of the very low-frequency band: (.003-.04 Hz)
LF power	ms^2	SNS and PSNS	Power of the low-frequency band: (.04-.15 Hz)
HF power	ms^2	PSNS	Power of the high-frequency band: (.15-.4 Hz)
Total power	ms^2	SNS and PSNS	Sum of the energy in the ULF, VLF, LF, and HF bands
LF/HF	–	SNS and PSNS	Ratio of LF to HF power
Non-Linear Methods			
SD1	ms		Poincaré plot standard deviation perpendicular to the line of identity
SD2	ms		Poincaré plot standard deviation along the line of identity
ApEn	–		Approximate entropy: measures the regularity and complexity of a time series
SampEn	–		Sample entropy: measures the regularity and complexity of a time series
DFA ($\alpha 1$)	–		Detrended fluctuation analysis: describes short-term fluctuations
DFA ($\alpha 2$)	–		Detrended fluctuation analysis: describes long-term fluctuations

Note: Originally published in Shaffer and Ginsberg (2017)
[1] NN is the RR interval after the signal preprocessing stage (filtered signal).
[2] SNS: sympathetic nervous system [3] PSNS: parasympathetic nervous system.
[*] There is a disagreement regarding the ANS contribution to ULF (Shaffer & Venner, 2013).

Frequency-Domain provides additional details about the frequency distribution of the HRV signal based on four bands: ultra-low frequency (ULF), very low frequency (VLF), low frequency (LF), and high frequency (HF). To obtain frequency measures, the time-series signal is transformed into the frequency domain by computing the power spectrum using a fast Fourier Transform (FFT) algorithm or autoregressive model.

Non-Linear Methods measure the unpredictability, irregularity, and complexity of the HRV signal. A Poincaré plot is a promising emerging technique that demonstrates changes in one RR interval as a function of the next RR interval. The plotted diagram presents information about any abnormalities detected in the HRV recording (Acharya et al., 2006). Thereafter, the standard deviation parameters (i.e., SD1 and SD2) can be derived from the Poincaré plot to identify correlations. In addition, the approximated entropy, sample entropy, and detrended fluctuation analysis (DFA) can be calculated to gain a deeper understanding of the non-linear behaviour of the HRV signal.

Given the complexity of the interactions among physiological systems, discerning which ANS branch influences each HRV measure has long been a matter of debate. Based on extensive research in the clinical and psychophysiological domain (Berntson et al., 1997; Malik et al., 1996), Shaffer and Venner (2013) summarised the dominant mechanisms for time- and frequency-domain HRV measures (see Table 2.1). Among the various HRV measures, RMSSD and HF power were most often used to reflect the vagal tone, which is associated with the PSNS; in contrast, SDNN reflects the activity of the SNS and PSNS (Laborde et al., 2017).

2.2.5 Ultra-Short-Term Analysis

Based on the standardised guidelines described by the Task Force (Malik et al., 1996), the most common durations for HRV analysis are long-term (24 h) and short-term (5 min). Long-term HRV recordings are mainly used in the assessment of pathological conditions related to cardiac disease and as a predictor of mortality risk following myocardial infarction (Kleiger et al., 1987; Kleiger et al., 2005). In contrast, short-term analysis is preferred for ambulatory HRV data acquisition given its practicality and ease of calculation. However, the existing literature has shown that short-term HRV analysis is not a strong predictor of mortality risk compared to long-term analysis (Malik et al., 1996). Nonetheless, short-term HRV is an effective measure of ANS dysfunction (Malik et al., 1996), particularly in psychophysiological studies (Alvares et al., 2016; Laborde et al., 2017; Quintana et al., 2016b).

More recently, there has been a surge of interest in the feasibility of analysing HRV in periods of less than 5 min to gain momentary insights regarding current psychophysiological states. As a result, UST analysis (< 5 min) has emerged in physical fitness and mental health contexts to measure HRV for eventual deployment in real-time applications and wearable devices (Shaffer et al., 2020). To reach these goals, researchers had to first assess the reliability and accuracy of UST HRV analysis compared to short-term HRV analysis (Burma et al., 2021; Pecchia et al., 2018; Shaffer et al., 2020). In general, the most common HRV measures selected for UST reliability evaluation were RMSSD, SDNN, HF power, and LF power, with the length of the investigated segments ranging from 10 s to 270 s. The majority of studies have focused on HRV data collected at a resting condition and analysed via statistical analysis (Baek et al., 2015; Esco & Flatt, 2014; Munoz et al., 2015; Nussinovitch et al., 2011), although a small number have examined the reliability of UST analysis in non-resting states, such as mental stress (Castaldo et al., 2019; Salahuddin et al., 2007b) and paced breathing (Melo et al., 2018).

However, there are compelling methodological concerns regarding the adequacy of the statistical tests commonly employed in UST HRV analysis. According to Shaffer et al. (2020), the statistical analyses utilised by Baek et al. (2015), Melo et al. (2018), Nussinovitch et al. (2011), and Salahuddin et al. (2007b) were not sufficient to assess the reliability of UST analysis. These analyses included group-mean differences and correlation tests, which do not provide information about the agreement levels between both measurements (UST vs. short-term). Hence, they recommended that agreement levels be assessed using Bland-Altman analysis (Altman & Bland, 1983), while acceptable measurement bias is taken into consideration. Such analyses have been performed by Castaldo et al. (2019), Esco and Flatt (2014), Munoz et al. (2015), and Shaffer et al. (2019).

This research seeks to address the paucity of the existing literature, particularly in terms of the evaluation of UST analysis in non-resting conditions (e.g., stress, paced breathing), by examining the reliability of UST analysis using limits of agreements supported by other statistical analyses (i.e., correlation and trend analysis) to facilitate the deployment of HRV in real-time applications and wearable devices (see Chapter 5).

2.2.6 Factors Affecting Heart Rate Variability

Several factors influence the measurement and analysis of HRV (Elliott & Moore, 2018; Fatisson et al., 2016; Sammito & Böckelmann, 2016), which are classified as follows:

- **Demographic factors,** such as age and gender, have a significant impact on HRV. The effect of age has been thoroughly documented via the inverse relationship between HRV and age due to the reduction of parasympathetic regulation with time (Agelink et al., 2001; Kuo et al., 1999; Voss et al., 2015). Regarding the effects of gender, women generally exhibited higher HRV measures compared to men, indicating that sympathetic activity is more dominant in men (Antelmi et al., 2004; Kuo et al., 1999; Voss et al., 2015).

- **Physical health** has a direct association with HRV. Specific physical health conditions that have a significant impact on HRV include cancer (Mouton et al., 2012), cardiovascular diseases (Haensel et al., 2008; Huikuri, 1995; Thayer et al., 2009b), diabetes (Benichou et al., 2018), respiratory diseases (Kazuma et al., 1997; Lutfi, 2012; Volterrani et al., 1994), and stroke (Binici et al., 2011; Lees et al., 2018). Research in this area has reported that HRV can be used as an indicator of physical health, given the positive relationship between them.
- **Mental health** has a clear effect on HRV similar to physical health. Psychiatric disorders, such as anxiety and depression, have been found to disrupt the balance of the ANS (Kemp & Quintana, 2013). Hence, recent research has demonstrated a link between low HRV levels and mental health issues. Additionally, HRV is influenced by cognitive workload (Luque-Casado et al., 2016), emotional reactions (Kop et al., 2011), and stress responses (Kim et al., 2018).
- **Physical activity** has a discernible impact on the ANS. Thus, HRV monitoring has been employed in athletic activities and sporting events to assess adaptability, endurance, fitness level, and training performance (Hamer & Steptoe, 2007; Plews et al., 2013). Moreover, HRV can predict the strength and power needed for athletes' recovery due to the increase

in sympathetic activity while performing physical exercise (Nakamura et al., 1993).

- **Nutrition** has been shown to have an impact on HRV levels. Both food type and nutrient amount provide the necessary building blocks for maintaining the digestion process, which activates the SNS and PSNS. Hence, low HRV values are linked to digestive disease and obesity (Strüven et al., 2021).

- **External factors** have been shown to affect HRV recordings (Quintana et al., 2016b). These factors include posture (i.e., sitting or supine), smoking status, and time of measurement (i.e., daytime or evening).

2.3 Biofeedback

The biofeedback discipline emerged from the intersection of psychology and medicine to treat mental and physical diseases. As the term implies, biological information is "fed back" to the user in order to regulate physiological activities and body functions (Brown, 1977). The definition of biofeedback was standardised in 2008 by the Association for Applied Psychophysiology and Biofeedback (AAPB), Biofeedback Certification International Alliance (BCIA), and International Society for Neuroregulation and Research (ISNR):

> Biofeedback is a process that enables an individual to learn how to change physiological activity for the purposes of improving health and performance. Precise instruments measure physiological activity such as brainwaves, heart function, breathing, muscle activity, and skin temperature. These instruments rapidly and accurately "feed back" information to the user. The presentation of this information – often in conjunction with changes in thinking, emotions, and behavior – supports desired physiological changes. Over time, these changes can endure without continued use of an instrument. (Schwartz et al., 2016, p. 17)

Ostensibly, there are two essential components of biofeedback systems: the instrument used to measure physiological signals and the mode of presentation used to convey the data collected via a display interface (e.g., audial, haptical, visual). A wide range of physiological biofeedback techniques (e.g., neurofeedback, respiratory biofeedback, thermal biofeedback) have been used in therapeutic settings for rehabilitation (Giggins et al., 2013) as well as the treatment of asthma (Lehrer et al., 2006), migraines (Nestoriuc & Martin, 2007), and psychiatric disorders (Schoenberg & David, 2014). Biofeedback is, thus, a promising non-invasive approach that can be used to improve health and well-being (Frank et al., 2010).

2.3.1 Heart Rate Variability Biofeedback

One of the most prevalent biofeedback mechanisms is the regulation of cardiac activity via the generation of RR sinusoidal oscillations produced by the interaction of multiple integrated regulatory systems, including autonomic, cardiovascular, and respiratory. Specifically, an increase or decrease in RR oscillations is facilitated by inhalation and exhalation, which activate the SNS and PSNS, respectively (Khazan, 2013). This phenomenon reflects the synchronisation of the cardiovascular and respiratory systems (see Section 2.2.2).

In HRVB, the ultimate aim is to maximise RR oscillations, which can be achieved using paced breathing exercises under a slow respiratory rate. The maximum HRV level, in which HR rhythms and breathing patterns are synchronised, is commonly referred to as the resonant frequency (RF). Normal RF rates for adults range between 4.5 to 6.5 breaths/min (Lehrer, 2007). Although Lehrer et al. (2000) proposed a protocol to determine a unique RF rate for each individual before commencing HRVB, recent studies have estimated that an RF rate of 6 breaths/min elicits similar physiological behaviour to a unique RF rate (Lehrer et al., 2020; Van Diest et al., 2014; Zaccaro et al., 2018). Breathing at this RF rate can achieve complete RSA synchrony, as reflected by a peak on the LF band of approximately 0.1 Hz (.04-.15 Hz). Additionally, the baroreflex contributes to RSA by sending feedback signals based on changes in BP sensed by the baroreceptors (Vaschillo et al., 2006). In simple terms, HR increases and BP decreases with inhalation; thus, the baroreflex causes an immediate increase in HR, and vice versa, to maintain homeostatic BP levels (Lehrer et al., 2003; Vaschillo et al., 2002). Overall, a combination of processes is involved in the HRVB mechanism during paced breathing at RF:

> The mechanism for this effect lies in a confluence of processes: (1) phase relationships between heart rate oscillations and breathing at specific frequencies, (2) phase relationships between heart rate and blood pressure oscillations at specific frequencies, (3) activity of the baroreflex, and (4) resonance characteristics of the cardiovascular system. (Lehrer & Gevirtz, 2014, p. 1)

2.3.2 Heart Rate Variability Biofeedback Techniques

Lehrer et al. (2000) first proposed a 10-session resonance breathing protocol to train individuals on HRVB techniques. Researchers have widely applied this protocol to investigate the long-term impact of HRVB on physical health, mental health, and cognitive performance (Lehrer et al., 2020). The long-term impact is assessed by comparing the HRV baseline measurement taken from the first session to the HRV measurement obtained from the last session. Nevertheless, Lehrer et al. (2013) simplified the training protocol and reduced the number of sessions to five for research and clinical purposes. In general, HRVB techniques involve paced breathing at approximately 6 breaths/min with prolonged exhalation (Lehrer & Gevirtz, 2014; Shaffer & Meehan, 2020). From a physiological perspective, exhalation stimulates the PSNS and significantly increases RSA, thereby increasing vagal tone.

To date, only a few studies have attempted to examine the short-term effects of HRVB interventions on vagal tone, focusing specifically on athletes (You et al., 2021a) as well as stress and other emotions (Laborde et al., 2022; Steffen et al., 2017; Wells et al., 2012). As an extension of these prior studies, Chapter 6 investigates the short-term impact of a single HRVB session on affective states, executive function, and physiological responses.

2.3.3 Feedback Modality

Although paced breathing practice without the incorporation of biofeedback elements has a positive influence on vagal tone (Laborde et al., 2019a; Laborde et al., 2021; You et al., 2021a), HRVB can provide the user with information that can potentially promote self-awareness about internal bodily processes and, by extension, improve self-regulation (Weerdmeester et al., 2020). In a systematic review of biofeedback approaches used to reduce mental stress, Yu et al. (2018b) found that HRVB is the most commonly used technique in stress management applications (19 out of 46), followed by multi-signal biofeedback (12) and respiratory biofeedback (8). The vast majority of the reviewed papers (35 out of 46) relied on visual displays as a means of representing biofeedback information to users, while the remaining studies used audio only (5) or a hybrid audiovisual approach (6).

More recently, research has introduced novel approaches to feedback modalities in the form of haptic feedback and virtual reality systems (Choi & Ishii, 2020; Yu et al., 2021), which have the potential to provide immersive environments that can enhance the HRVB experience (Blum et al., 2020; Rockstroh et al., 2019). Further, the incorporation of various feedback modalities in HRVB practices has the added benefit of increasing accessibility for individuals of all abilities.

2.4 Chapter Summary

This chapter offers an overview of the background details and pertinent research conducted in the context of affective computing, HRV, and HRVB as they relate to the primary research question motivating this thesis.

Given the physiological underpinnings of HRV, it has been regarded as a valuable indicator of both physical and mental health. The use of HRV to infer intrinsic affective states is a promising approach for recognising mental states in individuals, and the incorporation of HRVB practices can extend these benefits by improving individual mental well-being. With an aim to employ HRV in real-time affect recognition systems, four areas of limitation were identified in the existing literature:

- **HRV Preprocessing for real-time deployment**
  An essential requirement of reliable HRV analysis is the removal of artefacts to obtain a high-quality signal. However, two essential requirements of real-time HRV applications are flexibility and reliability. Prior studies have not been able to account for these two aspects of HRV acquisition and filtering for real-time analysis and batch processing (Benchekroun et al., 2021; Citi et al., 2012).

- **Signal length requirements for real-time HRV analysis**
  HRV is quantified using a sequence of RR intervals recorded across varying time periods; thus, the minimum reliable UST segment of the condition from which the data has been collected should be investigated prior to use in real-time HRV analysis applications. Recent studies have focused on the assessment of HRV measures in a resting state (Burma et al., 2021; Esco & Flatt, 2014; Shaffer et al., 2016). Moreover, a number of studies have drawn their conclusions based on an inadequate analytical

test rather than the assessment of the limits of agreements between the HRV measures derived from the UST segment and 5-min interval (Baek et al., 2015; Melo et al., 2018; Nussinovitch et al., 2011).

- **Short-term effects of HRVB on psychophysiological responses**
  The long-term effects of biofeedback activities on physical and mental health have been studied extensively (see Section 2.3). In particular, HRVB using paced breathing activities over multiple sessions can improve vagal tone, thereby improving a wide range of interconnected physiological responses and affective states (Lagos et al., 2008; Lee et al., 2015; van der Zwan et al., 2015). However, further research is needed to examine the short-term effects of HRVB practice on physiological responses and affective states.

- **Robust methodological implementations for affect recognition**
  Researchers have widely investigated the automatic detection of various affective states (e.g., emotion, stress) using physiological responses via ML algorithms. However, the use of limited datasets is prevalent within the field of affective computing, which raises challenges in system development and contextual performance interpretation (Castaldo et al., 2019; Foster et al., 2014; Schmidt et al., 2019). Thus, it is vital to ensure the adoption of robust methodological implementations to provide effective technological solutions to enhance mental health and well-being.

These limitations are addressed in the subsequent chapters by outlining the research methods used in this thesis (Chapter 3), followed by summaries of the four research studies conducted, each of which is detailed in its own chapter (Chapters 4-7).

CHAPTER 3

Research Methods

This chapter describes the methodological context underpinning the four research studies that comprise this thesis. It begins with a brief theoretical overview, followed by comprehensive details regarding the experimental design, data collection, data analysis, and ethical considerations.

3.1 Overview

Psychophysiological research has emerged as a potential means for investigating the relationship between physiological systems and psychological behaviour in human beings (Cacioppo et al., 2009). This research domain has been examined from an HCI perspective in a wide variety of applications by evaluating the psychophysiological responses under different conditions or developing effective interfaces for advanced monitoring (Cowley et al., 2016). Broadly, HCI researchers have used qualitative, quantitative, and mixed research designs to explore and identify affective states through physiological responses (e.g., HR, skin conductance). This thesis adopts a hypothetico-deductive model for quantitative data collection and analysis guided by the following main research question:

> **RQ:** How does a single HRVB session using paced breathing mediate physiological responses across a range of affective states, and can these affects be robustly recognised by supervised learning algorithms?

The hypothetico-deductive model is a scientific approach that focuses on formulating falsifiable hypotheses to assess theories or observations through empirical studies (Hayes, 2021). The primary goal of this research was to test multiple hypotheses to gain further insight into the relationship between increased levels of HRV and mental well-being, as indicated by psychophysiological responses (i.e., affective states, executive function, and physiological measures) to develop real-time affect recognition systems (see Figure 3.1). The employed research paradigm aligns with a positivist epistemology, which focuses on constructing knowledge through experimental research, quantified measurements, and statistical analysis (Hayes, 2021).

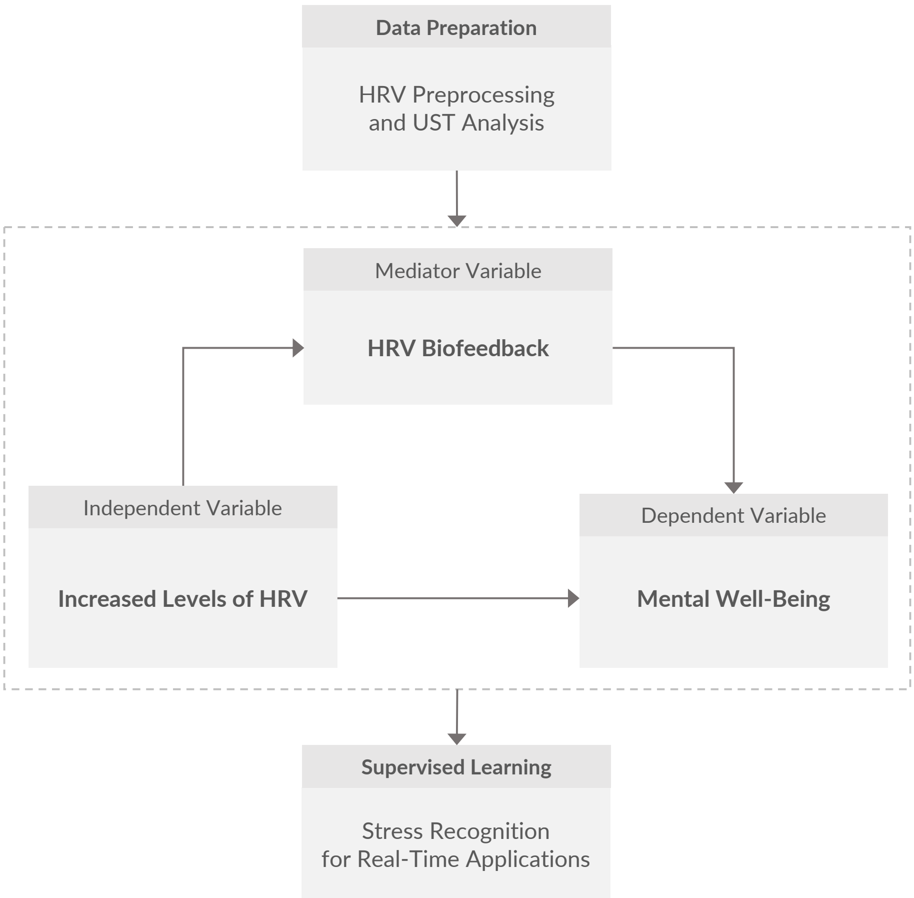


Figure 3.1: Conceptual Framework of the Research Approach

The following sections provide comprehensive details about the experimental design, data collection, data analysis, and ethical considerations of the four studies comprising this research.

3.2 Experimental Design

Chapters 4-7 describe the four research studies of this thesis, each of which is led by a distinct sub-RQ (see Section 1.2). The details of the study designs are as follows:

CH4: HRV preprocessing study — a reliability analysis study was conducted to filter HRV data from publicly available datasets to facilitate batch processing and real-time analysis. The filtering process involved artefact detection and correction methods; thus, common methods were examined and compared against the Kubios results to test for the relationship and agreement levels. Further, an open-source framework to communicate with HRV sensors over Bluetooth was developed and integrated with the preprocessing algorithms to enable HRV data acquisition and filtering in real time.

CH5: UST exploratory study — a repeated measures quasi-experiment was conducted using a three-level independent variable: baseline, stress, and paced breathing. The dependent variable (HRV) was measured for all participants under these three different conditions. Based on the documented physiological relationship of HRV with stress conditions and paced breathing exercises (see Sections 2.1.1 and 2.3.1), these two stimuli (i.e., stress and paced breathing) were chosen to evaluate the consistency of HRV within the investigated UST segments. All participants experienced the same conditions in the same order to facilitate the comparison of HRV during paced breathing to baseline after going through a stress phase. As a result, there was no randomisation in the condition assignment, hence the quasi-experiment design. A standardised procedure was followed to ensure that all participants had the same experience, thereby reducing the potential effect of confounds (e.g., location, time of HRV measurement).

CH6: HRV biofeedback study — a randomised controlled trial (RCT) based on a mixed-factorial design was conducted with two groups: intervention group performing HRVB through paced breathing and a control group (CTRL) breathing at a normal rate. The dependent variables analysed in this study were affective states, cognitive performance (working memory task), and physiological measures (HRV and BP). Self-reported questionnaires focused on various affects, such as perceived stress and relaxation levels, were used to assess affective states (see Section 3.3.2). The HRV data and self-reported questionnaires were collected four times: at baseline, pre-, mid-, and post-intervention. The cognitive task took place pre- and post-intervention. BP was measured at baseline, pre-, and post-intervention to examine BP changes in response to the cognitive stress task. The independent variables were group (i.e., CTRL or HRVB; between-subjects) and time (i.e., baseline, pre-, mid-, post-intervention; within-subject). Group assignment was randomised using a random number generator. Similar to the previous study, a standardised procedure was followed to minimise confounding effects.

CH7: Stress recognition study — a binary classification study using supervised learning algorithms was carried out to recognise stress and relaxation levels from HRV data. The classifiers were developed using robust ML strategies addressing limitations related to data segmentation, feature selection, and model evaluation. To assess for generalisability, the data discussed in Chapters 5 and 6 were used for training, and public independent datasets were used for testing.

3.3 Data Collection

Two main sources were incorporated for data collection: primary and secondary. The primary sources comprised data acquired from participants, and the secondary sources comprised data acquired from online public datasets. Several types of data were collected as primary sources, such as psychological data via questionnaires, physiological data via sensors, and task performance data via computer-logged measurements.

3.3.1 Participants

The characteristics of the targeted population were predetermined as it was not practically feasible, given the available time and resources, to perform experiments on the entire population or a genuinely random sample. Samples of the population were drawn based on the following characteristics: healthy individuals between 18 and 65 years of age with no physical health conditions (e.g., cardiovascular disease, respiratory disease) and no severe mental disorders (e.g., dementia, depression). Power analysis, effect size, and previous research helped to determine the target sample size required to maintain a minimum power of 80%. Statistical power analysis for each study is discussed in more detail in the relevant chapters.

For the UST exploratory study discussed in Chapter 5, participants were recruited from Queen Mary University of London (QMUL) via a call for participation sent to students and faculty in the computer science department. For the HRV biofeedback study discussed in Chapter 6, an email announcement was distributed to the mailing list of Hamad bin Khalifa University's (HBKU) in the State of Qatar. Personal invitations were also sent by email to friends and family members who expressed interest in the study. Participants in all studies were fully informed about the nature of the experiment and signed consent forms.

Due to the COVID-19 pandemic and subsequent lockdowns in the United Kingdom, it was not possible to conduct face-to-face experiments for HRV data collection. Therefore, the HRV biofeedback study was conducted in a collaboration with HBKU in the State of Qatar through the Qatar National Research Fund (QNRF). The principal investigator of the study, Dr Dena Althani from HBKU, reviewed the research study application and assisted in the data collection process. My contributions included study conception and design, data collection, data analysis, and interpretation of results. The study was conducted in compliance with the Qatar government's precautionary guidelines to limit the spread of the infection.

3.3.2 Questionnaires

Several questionnaires were employed to collect demographic information about participants as well as factual information about their feelings and behaviours. These questionnaires were used in the studies that involved data collection from participants (Chapters 5 and 6) unless stated otherwise. Refer to Appendix A for the full questionnaires.

Demographic attributes, such as age, body mass index (BMI), and gender, are relevant factors in the assessment of cardiac activity. Hence, a demographic questionnaire was completed by all participants at the beginning of the study to gain a better understanding of their backgrounds. Moreover, an HRV-related questionnaire was applied as a pre-screening survey and to establish a baseline for comparable results among participants, as recommended by Quintana et al. (2016b). In general, the questions were designed to identify participants who should be excluded from the study due to such factors as cardiovascular disease, mental disorder, sleeping patterns, and smoking within the last 24 hours before the experiment.

Mental health states were analysed using self-reported questionnaires addressing multiple affective states. Chapter 5 reports the results of the Generalized Anxiety Disorder Scale (GAD-7), which consists of seven self-reported items focused on symptom severity over the last two weeks. The questionnaire has excellent internal consistency, as indicated by a Cronbach's alpha of .92, and good test-retest reliability, as indicated by an intraclass correlation of .83 (Spitzer et al., 2006).

Focusing on a range of affective states, Chapter 6 reports the results of the Depression Anxiety Stress Scale (DASS-21; Lovibond & Lovibond, 1995) as well as the Positive and Negative Affect Schedule (PANAS) questionnaire (Watson & Clark, 1994). Each subscale of the DASS-21 consists of seven self-reported items that ask about the extent to which statements apply to the individual over the last month. According to Antony et al. (1998), the DASS-21 demonstrates a high level of internal consistency for depression, anxiety, and stress, as indicated by Cronbach's alphas of .94, .87, and .91, respectively.

In comparison, the PANAS assesses both positive and negative affective states over the last week using two 10-item scales, one for positive affect and one for negative. The internal consistencies for positive and negative affect are moderately good, as indicated by Cronbach's alphas with a value greater than .83 for both (Watson & Clark, 1994). In addition, three components of the expanded version of the PANAS (PANAS-X) were used to assess attentiveness, fatigue, and serenity during the study, with questions focused on the participant's feelings at the moment. Finally, five-point single-item Likert scales were used to assess general stress and mood levels at different time points during the study.

In addition to affective states, Chapter 6 addresses sleep quality using the Pittsburgh Sleep Quality Index (PSQI), a self-reported questionnaire that assesses seven components: sleep quality, sleep duration, sleep latency, sleep efficiency, sleep disturbances, use of sleep medication, and daytime dysfunction over the previous month. The questionnaire has moderately good internal consistency, as evidenced by a Cronbach's alpha of .83 (Buysse et al., 1989). Further, Chapter 6 considers physical activity via the short-form version of the International Physical Activity Questionnaire (IPAQ), which estimates total physical activity and total time spent sitting over the previous week using seven self-reported items. The questionnaire has moderately good consistency, as indicated by a test-retest Spearman's reliability of .80 (Craig et al., 2003).

All questionnaires used in Chapter 6 were presented in the English and Arabic languages as most of the participants were non-native English speakers. The Arabic versions of the questionnaires were obtained as follows: PANAS (Davis et al., 2020), DASS-21 (Ali et al., 2017), PSQI (Suleiman et al., 2010), and IPAQ (Helou et al., 2017). The remaining demographic and HRV-related questionnaires were translated by the researcher (see Appendix E).

3.3.3 Physiological Data

Information about the intrinsic behaviour of the human body can be attained by collecting data through physiological sensors. However, challenges may arise in experimental research studies that implement physiological assessment, particularly in terms of data collection, configuration, analysis, and interpretation (Lazar et al., 2017). Therefore, sensors were selected based on the reliability of the signal as well as the ease and convenience of their attachment to the participant. Moreover, the studies were conducted in a laboratory setting to provide a comfortable and controlled environment while minimising confounding variables.

To capture cardiac activity under different experimental conditions, HRV was measured using the CorSense device by Elite HRV[1], which generates a PPG signal recorded at a sampling rate of 500 Hz (see Figure B.1). Recent research has shown that PPG can provide accurate HRV measures that correlate closely with ECG-derived measures for healthy subjects. Moreover, CorSense can be easily attached to the user's finger, ensuring comfort and convenience compared to the conventional chest strap (Pasadyn et al., 2019).

To determine device reliability, an intraclass correlation coefficient (ICC) with a 95% confidence interval (CI) was used to compare the CorSense PPG signal to both a three-leads ECG signal obtained using a BIOPAC MP150 device and a PPG signal obtained using a Polar H7 chest strap sensor; readings were taken from a single participant under normal breathing conditions. The correlation coefficients were .84 (95% CI [.82, .86]) with BIOPAC and .85 (95% CI [.82, 87]) with Polar.

The raw RR intervals of CorSense were exported offline (i.e., after data collection) using an iOS application developed by the sensor's manufacturer (Elite HRV). In general, CorSense transmits data over Bluetooth 4.0 using the Bluetooth Low Energy (BLE) protocol with a Bluetooth-enabled device (e.g., smartphone). To obtain raw data in real-time and integrate CorSense with various research-based applications, an interface using the Bluetooth standards was developed as part of the study discussed in Chapter 4.

All signals were subjected to a visual inspection to ensure that they conformed to the expected norms and did not contain any unanticipated trends or outliers. Kubios software was primarily used for the HRV analysis in the study described in Chapter 4, and for the HRV data exploration in the studies described in Chapters 5-7; however, the HRV analysis was performed in Python

[1] https://elitehrv.com/corsense

using the Systole (Legrand & Allen, 2022) and pyHRV (Gomes, 2018) packages. In the HRV biofeedback study, systolic blood pressure (SBP) and diastolic blood pressure (DBP) were recorded using an OMRON M7 Intelli IT upper arm cuff monitor.

External factors that can affect cardiac activity measurement were considered. Accordingly, participants were asked to avoid alcohol, coffee, heavy meals, and intensive workouts prior to the experiment.

3.3.4 Cognitive Performance

There is a well-established relationship between executive function and vagal tone (Hansen et al., 2003; Kim & Lee, 2013; Thayer et al., 2009a). Working memory is a facet of executive function characterised by the capacity to retain limited information for a short period of time while mentally processing the information (see Section 2.1.2; Baddeley & Hitch, 1974). The study described in Chapter 6 investigated the effects of HRVB on executive function and, by extension, working memory, as measured by the N-back cognitive task (Kirchner, 1958). Cognitive performance was determined by calculating the correct responses and reaction time. Reaction time was measured as the time required for the participant to make a decision about the correct response.

3.3.5 Datasets

Although there were several publicly available datasets containing ECG and HRV data, careful consideration was given to selecting the appropriate dataset based on the study requirements. Following a review of the datasets concerning the experiment condition, number of participants, signal length, signal quality, and study protocol, one dataset was selected for Chapter 4, and two datasets were selected for Chapter 7 (see Table B.1).

1. **Normal breathing group**

   This group consisted of ECG data for 11 healthy participants (eight women, three men; ages 20-35) breathing spontaneously during sleep with an average duration of 6 h each (Peng et al., 1999). Data were taken from the publicly available PhysioNet database (Goldberger et al., 2000). This dataset was used to assess the performance of artefact detection and reliability of correction methods because the HRV data were collected from healthy participants in a stable condition and the dataset included annotation files confirming normal heartbeats. The HRV data of the normal breathing group has been used in relevant research studies (Deka & Deka, 2020; Gonzalez et al., 2012; Muñoz Diosdado et al., 2010). Further detail about the dataset is discussed in Section 4.3.1.

2. **WESAD**

   Wearable Stress and Affect Detection Dataset (WESAD) is a publicly available multimodal dataset consisting of physiological data recordings, including body temperature and three-axis acceleration, ECG, electrodermal activity, electromyograms, and respiration recorded during baseline, stress, meditation, and amusement conditions using a chest belt and wrist sensors. Data were collected from 15 participants in a controlled laboratory experiment and physiological signals were sampled at 700 Hz (Schmidt et al., 2018). In addition, self-report surveys were administered to gauge stress and emotional states. This dataset has been widely used in relevant research studies (Chakraborty et al., 2019; Elzeiny & Qaraqe, 2020; Jiang et al., 2020; Sarkar & Etemad, 2020). All conditions except for the data collected during the amusement phase were employed in the present study. WESAD was used as an independent dataset to assess the generalisability of the supervised learning algorithms (see Section 7.4.1).

3. **SWELL**

   Smart Reasoning Systems for Well-being at Home and at Work (SWELL) is a publicly available dataset collected by researchers at the Institute for Computing and Information Sciences at Radboud University (Koldijk et al., 2014). It consists of computer recordings of body posture, ECG signals, facial expressions, and skin conductance from 25 participants performing two work-related tasks under two types of stress induction (i.e., receiving unexpected email interruptions and pressure to complete their work within a certain timeframe). ECG signals were sampled at 2048 Hz. In addition, the researchers collected subjective information regarding the participants' emotions, mental effort, perceived stress, and task load. This dataset has been widely used in relevant research studies (Behinaein et al., 2020; Koldijk et al., 2018; Sarkar & Etemad, 2020; Sriramprakash et al., 2017). Like WESAD, SWELL was used as an independent dataset to determine the generalisability of the supervised learning algorithms discussed in Section 7.4.1.

3.4 Data Analysis

Overall, the collected data were quantitatively analysed based on descriptive statistics, inferential analysis, and predictive analysis.

First, a general overview of the collected data was collated using descriptive statistics; the central tendency, dispersion measures, and CI were reported for each variable as appropriate. Moreover, correlation analyses were employed to evaluate the relationship between variables and their agreements: specifically, Pearson correlation coefficient, Spearman's rank-order correlation, ICC, and Bland-Altman analysis. The latter is a graphical approach that calculates the 95% limits of agreement and plots the mean of both measurements against differences in means (Altman & Bland, 1983; Giavarina, 2015). As a quantification measure for the Bland-Altman analysis, the bias is calculated as the mean difference of both measurements.

Second, inferential statistical methods were used to gain deeper insight into the data and assess whether the observed differences between groups were reliable or merely coincidental. After summarising the variables through descriptive statistics, the data were checked for normality using the Shapiro-Wilk test and for homogeneity of variance using Levene's test; an appropriate statistical test was selected accordingly (i.e., parametric or non-parametric). Moreover, the sphericity assumption in repeated measures analysis was assessed using Mauchly's sphericity test. To control for Type I errors, the significance levels were set at 5% for all statistical analyses. To achieve practical significance, the effect sizes were reported and both sample size and statistical power were taken into consideration, minimising the likelihood of Type II errors based on Cohen's standards (Cairns, 2019; Cohen, 1988).

The group-mean differences statistical tests used in this research were derived from a broad family of statistical models known as the generalised linear model (GLM), including analysis of variance (ANOVA), analysis of covariance (ANCOVA), multilevel linear (MLL) analysis, and t-tests. In addition, linear and logistic regressions were used to perform predictive analysis. Finally, ML models using supervised learning algorithms were developed to recognise stress and relaxation levels based on binary classifiers. The statistical analyses were performed with RStudio (version 4.0.0), and the ML models were implemented in Python using the Scikit-Learn package (Barupal & Fiehn, 2019).

Table 3.1 presents a comparative summary of the experimental studies based on the methodological attributes of each study.

Table 3.1: Summary of the Experimental Studies

	Study 1 Chapter 4	Study 2 Chapter 5	Study 3 Chapter 6	Study 4 Chapter 7
Study Name	HRV Preprocessing	UST Exploratory	HRV Biofeedback	Stress Recognition
Study Objective	Develop reliable and flexible implementation for HRV filtering	Explore the minimum reliable segment for HRV analysis	Investigate the short-term effects of HRVB on psycho-physiological measures	Evaluate a robust implementation for stress recognition using HRV data
Experiment Design	–	Quasi-Experiment	Mixed-Factorial Design	–
Data Analysis	Correlation and Bland-Altman	Correlation, Bland-Altman, and MLL	ANCOVA, Correlation, MLL, and Regression Analysis	Supervised Learning Algorithms
Data Source	Secondary (Public Dataset)	Primary Data Collection	Primary Data Collection	Primary and Secondary: Dataset of Studies 2 & 3 and two Public Dataset
Participants Number	11	20	38	96

3.5 Ethical Considerations

Institutional approval was obtained from the Ethics Committee at QMUL and the Institutional Review Board (IRB) at Qatar Biomedical Research Institute at HBKU, as appropriate. Participation was voluntary, and all participants had the opportunity to read the study details on an information sheet and were asked to sign a consent form (forms provided by QMUL and HBKU). Accordingly, participants were aware of their right to withdraw at any time if they desired, with no adverse consequences. Additionally, a debriefing session was held for participants interested in obtaining further details about the study's objectives and preliminary results.

Several strategies were employed to maintain data anonymity and confidentiality. First, participants were identified using a randomly generated participant number and asked not to include any identifying information in their questionnaire responses (e.g., name, mobile number, email address). Second, a unique coding scheme (XXXX-XXXX-XXX) was used to store all data files for each participant in three parts: 1) data type (e.g., questionnaire or physiological measure), 2) group association, and 3) participant ID. Lastly, the data were stored in a password-protected hard drive only accessible to the researcher conducting the experiment.

3.6 Chapter Summary

This chapter presents an overview of the research methods and experiment designs used in the four primary studies. The research approach implemented required several data collection methods: namely, physiological data, psychological data, and task performance data, which are discussed at greater length in Chapters 5 and 6. Further, public datasets, discussed in Chapters 4 and 7, were employed to provide data diversity and enrichment through the incorporation of secondary sources. The data used in Chapters 4-6 were analysed quantitatively using statistical analyses, and the data used in Chapter 7 were analysed using supervised learning algorithms. In-depth details about the methods and procedures for each study are explicated in the methods section of the relevant chapters. Lastly, ethical considerations for all studies involving participants are also discussed.

CHAPTER 4

Heart Rate Variability Preprocessing

This chapter serves as the foundation for the subsequent three chapters by providing methods to obtain high-quality HRV data, with particular attention to the utilisation of preprocessing algorithms (artefact detection and correction) that can facilitate batch processing and real-time analysis for eventual deployment in wearable devices. Lastly, an open-source framework for real-time HRV data acquisition using Bluetooth-based sensors is described.

4.1 Overview

The reliability of HRV analysis depends, to a great extent, on the signal quality of the RR intervals. Hence, a preprocessing step is imperative to identify and correct any existing artefacts. Typically, the presence of an artefact is due to either physiological reasons or technical issues in the HRV data collection. A physiological artefact appears when the heart produces abnormal beats, such as ectopic beats or ventricular fibrillation (Logier et al., 2004; Peltola, 2012). In contrast, technical artefacts are caused by the physical motion of the sensor or the user (Peltola, 2012).

Traditionally, trained investigators would visually inspect ECG signals to identify abnormalities in the QRS complex, after which abnormal beats and other artefacts were edited and corrected offline (Soler et al., 2018). However, this process has been adapted to efficiently manage long-term HRV data and large datasets via automatic artefact detection and correction, thereby enhancing batch processing. An additional advantage of automatic filtering is that it supports HRV data collection via wearable sensors by meeting the requirements of real-time digital health applications. Thus, various algorithms have been proposed to perform automatic filtering of HRV signals (Jarrin et al., 2012; Lang, 2019; Ribeiro et al., 2018), which comprises two major stages: 1) artefact detection and 2) artefact correction.

Among clinicians and researchers, Kubios is one of the most widely used commercial applications for reliable HRV filtering and analysis (Tarvainen et al., 2014). The standard version of Kubios is a freeware HRV analysis application that provides five predefined levels of artefact detection based on a fixed-threshold approach (Tarvainen et al., 2016). To detect artefacts, each RR interval is compared to a local average value obtained via median filtering. If the difference between the RR interval and local average exceeds the predetermined threshold level, the RR interval is marked as an artefact. Afterwards, the identified artefacts are corrected using cubic spline interpolation. However, while these commercial applications are both reliable and efficient, they tend to be closed source, which places considerable limitations on the performance of batch processing and real-time analysis.

Accordingly, this study sought to investigate the following research question by developing essential detection and correction algorithms to flexibly serve the requirements of batch processing and real-time analysis while maintaining high agreement levels with Kubios results:

> **SRQ1:** What signal preprocessing algorithms are necessary for a reliable real-time HRV analysis?

Hence, the open-source Python programming language was used to import raw HRV data as well as identify and correct artefacts based on the recommended filtering methods from the literature discussed in Section 4.2. In addition, an interface for CorSense was implemented using Python to enable real-time HRV data collection via Bluetooth technology.

4.2 Related Work

4.2.1 Detection Techniques

Having an automatic strategy for artefact detection in HRV signals can improve analysis efficacy and, by extension, promote the use of HRV in clinical and academic research. There are two main approaches for identifying existing signal artefacts: 1) fixed threshold and 2) adaptive threshold. See Section 2.2.3 for an overview of artefact types and signal preprocessing techniques.

Fixed Threshold relies on the assignment of a specific threshold to determine whether the RR value is a potential artefact. Previous studies have based their artefact detection strategies on Malik et al. (1989), which established that there will be an abrupt change in the signal indicating the presence of an artefact when the difference between each consecutive RR value (dRR; see Equation 4.1) is greater than a specific value (usually 20%; **absolute fixed**):

$$dRR(i) = |\, RR(i) - RR(i-1) \,|, \quad i = 2, 3, ..N \tag{4.1}$$

Recently, a variety of flexible methods have been developed to determine whether an RR interval is an artefact by comparing the RR interval to a local average or median (mRR; see Equation 4.2) and calculating whether the difference exceeds a predefined threshold (**median fixed**). This approach is implemented in the standard Kubios application using the following threshold levels: 1) very low (450 ms), 2) low (350 ms), 3) medium (250 ms), 4) strong (150 ms), and 5) very strong (50 ms; Tarvainen et al., 2016).

$$mRR(i) = RR(i) - median\,[\,RR(i-5, ..., i+5)\,]\,, \quad i = 1, 2, ..N \tag{4.2}$$

Adaptive Threshold extends the median-fixed approach and replaces the predefined threshold with a varying threshold. Lipponen and Tarvainen (2019) proposed an adaptive threshold that is dependent on the time-varying sequence of RR intervals; in other words, the threshold value changes based on the distribution of the median RR intervals. First, the algorithm computes the mRR by considering the local median value of 10 RR intervals surrounding the investigated RR value, as shown in Equation 4.2. Second, the adaptive threshold variable (θ) is calculated using the quartile deviation (QD) of the differences in the mRR based on a window consisting of 90 RR intervals. This threshold variable is illustrated in Equation 4.3, where α is a constant of 5.2, as recommended by Lipponen and Tarvainen (2019). The QD is defined as the product of half the difference between the first and third quartiles. Finally, if mRR is greater than the adaptive threshold (θ), then the RR value is identified as an artefact.

$$\theta(i) = \alpha\, QD\,[\,|\,mRR(i-45, ..., i+45)\,|\,] \qquad i = 1, 2, ..N \tag{4.3}$$

4.2.2 Correction Techniques

After identifying the location of the artefacts, the appropriate correction method can be applied. The most common methods for artefact correction are briefly summarised as follows (Peltola, 2012):

1. **Deletion** removes the identified artefacts from the original signal and shifts the subsequent RR intervals, resulting in a shorter signal.

2. **Window Average** replaces the identified artefacts with the mean or median of the neighbouring RR values. The median method can be specifically used to minimise the effect of any other outliers (Thuraisingham, 2006), as shown in Equation 4.4 below for signal $s(n)$:

$$s_i'(n) = median\left[s(n+m) : |m| \leqslant \frac{w_m - 1}{2}\right] \tag{4.4}$$

Note. where $s_i(n)$ is the averaged signal at time n and w_m is the length of the window centred around n.

3. **Interpolation** replaces the artefact with a value calculated from the surrounding data by fitting a straight line (linear interpolation) or smooth curve (cubic spline interpolation), the latter of which is estimated from a cubic polynomial.

4.3 Methods

4.3.1 Dataset

The normal breathing group dataset (Peng et al., 1999) from Physiobank (Goldberger et al., 2000) was used to assess performance and reliability in the filtering process. The dataset included a set of ECG recordings with annotations collected from 11 healthy participants (eight women, three men; ages 20-35 years) breathing spontaneously during sleep for an average duration of 6 h each. Although HRV measures may vary depending on state (e.g., sleep, conscious) and seating position (e.g., supine, seated), the goal of this study was to obtain stable HRV data from healthy participants with a low number of artefacts in order to synthetically introduce the artefact type of interest.

The RR intervals were computed based on the time difference between the R peaks using the Waveform Database Python library provided by PhysioToolkit (Goldberger et al., 2000). Subsequently, four short-term segments were extracted from each participant manually for 6 min, resulting in 44 RR segments. Although all R peaks were labelled as normal beats, the data were visually inspected to select segments free from extreme outliers.

4.3.2 Artefacts Simulation

In accordance with the methods employed in the previous research (Citi et al., 2012; Lipponen & Tarvainen, 2019), the artefacts were then simulated by synthetically adding long RR intervals to the signal at different indices of $i = 10n$, where $n = \{1, 2, 3..\}$. The addition of the long intervals emulated the presence of technical artefacts introduced by sensor movements as well as missed beats (see Figure 2.3; Peltola, 2012). The goal was to add a maximum of 10% artefacts with respect to the total number of samples. In total, there were 20,851 samples, of which 1,718 were simulated artefacts. Subsequently, RR signals with simulated artefacts were manually imported into the Kubios software, and three filtering thresholds were applied (low, medium, and strong). The medium threshold was selected as it obtained the lowest artefact detection error rate (2.6%) of the three threshold levels (low: 25%, strong: 128%). Finally, the resulting filtered signal obtained from Kubios was used as the reference for analysis in the current study. The following naming conventions were given to the different signals used:

- **Original signal** – extracted RR intervals from the ECG signal with a 6-min duration.
- **Erroneous signal** – RR intervals of the simulated artefacts.

- **KUB-corrected signal** – RR intervals filtered using the Kubios filtering method configured with a medium threshold. This signal was used as a reference for the reliability analysis of the correction methods.
- **KUB-detected signal** – RR intervals of the detected artefacts using the Kubios filtering method. These were generated by marking artefact indices resulting from the subtraction of the erroneous signal from the KUB-corrected signal. This signal was used as a reference for the performance analysis of the detection methods.

4.3.3 Artefact Detection

Procedure

The implemented artefact detection process involved three threshold-based techniques: 1) absolute-fixed threshold, 2) median-fixed threshold, and 3) adaptive threshold. All techniques were implemented in Python using standard libraries, such as pandas (McKinney, 2010) and NumPy (Harris et al., 2020). The source code for the three detection algorithms is available in the researcher's repository on Github (Bahameish, 2019).

As described in Section 4.2.1, the absolute-fixed approach was performed by calculating the difference between consecutive RR values, then assessing whether the absolute difference was greater than 20% of its preceding RR value. In contrast, the median-fixed threshold algorithm was similar to the detection method used in Kubios. First, a local median value was calculated for a moving window with a size of 10 RR values before comparing the difference between the median and individual RR values against a threshold of 250 ms, which corresponded to Kubios' medium threshold. Following the procedures recommended by Lipponen and Tarvainen (2019), the adaptive threshold was based on the time-varying threshold of a series of median RR intervals (see Equations 4.2 and 4.3).

Algorithm 1 provides a pseudocode of artefact detection methods to elucidate the threshold calculation for each approach. Figure 4.1 depicts the detection results for five synthetically inserted long RR intervals: the signal with artefacts is represented by a black line, dRR and mRR are represented by a blue line, and the computed threshold for each method is represented by a dashed line. Out of five artefacts, the absolute-fixed approach missed two, and the median-fixed approach missed one. In contrast, the adaptive threshold approach identified all artefacts correctly in addition to marking one extra RR interval as an artefact.

Algorithm 1: Artefact Detection Algorithms

```
# Initialisation: wsm = 11, wst = 91, α = 5.2.
# wsm: Window Size for mRR
# wst: Window Size for adaptive threshold
function detectArtefacts(RR, method, wsm, wst):
    # Refer to Equations 4.1-4.3 for details about i and N.
    dRR_i = abs( RR_i − RR_{i−1} )
    mRR_i = RR_i − median [ RR(i − wsm/2, ..., i + wsm/2) ]
    θ(i) = α QD [ | mRR(i − wst/2, ..., i + wst/2) | ]
    # Get artefact indices based on the selected method.
    if method = absolute then
        index = get_index(dRR_i > 0.2RR_i)
    else if method = median then
        index = get_index(mRR_i > 250)
    else if method = adaptive then
        index = get_index(mRR_i > θ_i)
    return index
```

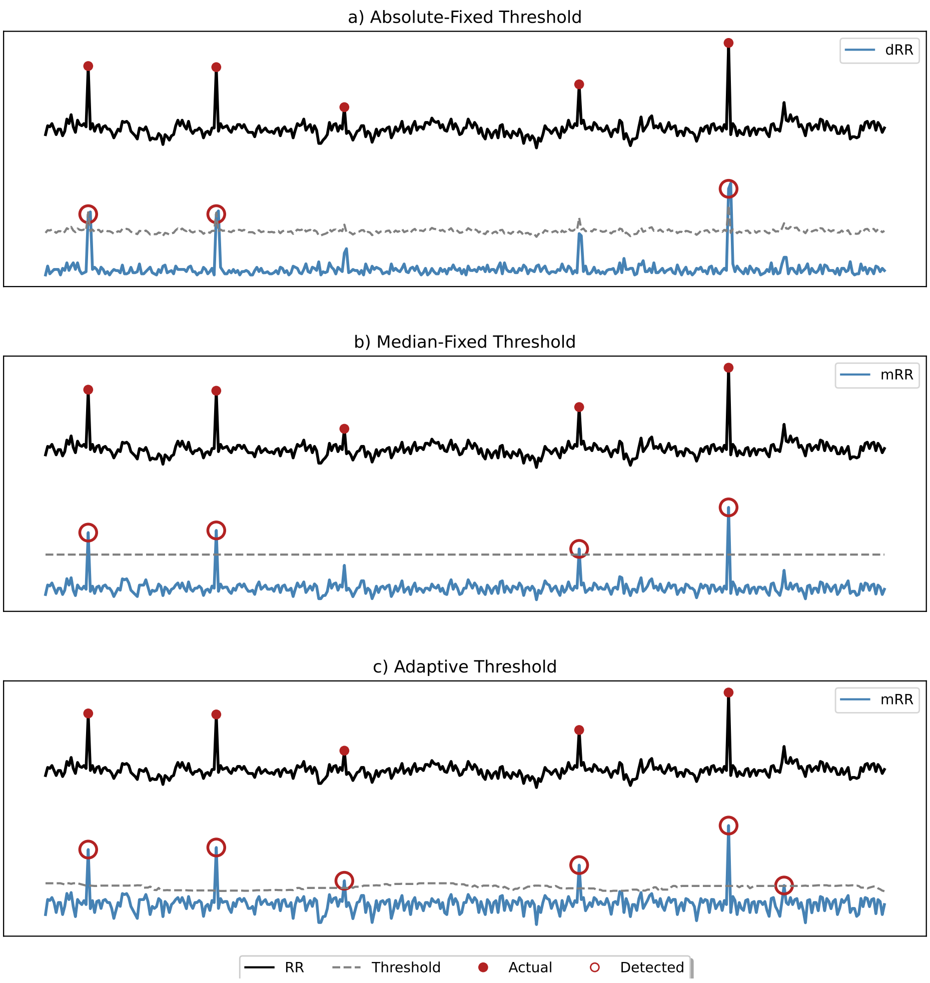


Figure 4.1: An example of the detection algorithms for five simulated artefacts.

Performance Metrics

To evaluate the detection method performance, the confusion matrix approach was employed to classify each RR interval as a normal or abnormal instance while using the Kubios results as a reference. Accuracy, balanced accuracy, sensitivity, and specificity were also computed as performance metrics. Lastly, the processing time was calculated for each detection method. The performance metrics of this study were defined as follows:

Accuracy measures the ratio of instances correctly classified as normal or abnormal to the total number of instances in each class.

Balanced accuracy measures the average class accuracy by calculating the arithmetic mean of sensitivity and specificity metrics. It is often used for imbalanced datasets (e.g., anomaly detection).

Sensitivity measures the ratio of instances correctly classified as abnormal to the total number of abnormal instances in the dataset.

Specificity measures the ratio of instances correctly classified as normal to the total number of normal instances in the dataset.

Processing time measures the average time taken for the algorithm to identify artefacts in the RR intervals.

In this study, balanced accuracy was selected as the main performance metric for accuracy assessment due to class imbalance (i.e., normal and abnormal instances).

Although Kubios provides the total number of detected artefacts, it does not provide information about their indices. Hence, a simple script was developed to identify the indices by importing the reports generated by Kubios, extracting the RR intervals from the KUB-corrected signal, and comparing these intervals to their corresponding values in the erroneous signal. As a result, if the difference between both intervals was not zero, the index was marked as an artefact identified by Kubios and saved as a new signal (KUB-detected). Kubios detected 1,763 artefacts in total, which was slightly higher than the number of simulated artefacts in the erroneous signal (1,718).

4.3.4 Artefact Correction

Procedure

After the artefact identification, the RR signals were corrected based on the techniques recommended by Peltola (2012). Three methods were assessed in this study: 1) deletion, 2) moving window average, and 3) cubic spline interpolation. Deletion entailed the removal of an identified artefact sample from the signal. Moving window average involved calculating a new RR value based on the mean of 10 samples surrounding the identified artefact. Cubic spline interpolation replaced the artefact by fitting a third-degree polynomial from 10 samples to estimate a smooth curve. Subsequently, all corrected signals were analysed using Kubios to compare the reliability of the resulting HRV measures. For each correction method, the Kubios application generated 44 reports. Hence, the developed script was used to automate the processes of importing all Kubios reports and extracting the computed HRV measure for reliability analysis.

Reliability Analysis

An ICC with a 95% CI was used to assess the reliability of the correlation and agreement levels between the HRV measures derived from the three correction methods and KUB-corrected signals. The ICC was computed using a two-way mixed-effects model based on a single measurement and absolute agreement.

The results were supported by a Bland-Altman analysis (Altman & Bland, 1983), which was performed by calculating the 95% limits of agreement and plotting the mean of both measurements against the difference between them. In this study, the strength of the ICC reliability level was interpreted as follows: 0-.50 poor, .50-.75 moderate, .75-.90 good, and .90-1 excellent (Koo & Li, 2016). All HRV measures were log-transformed due to violation of the normality assumption, as assessed with the Shapiro-Wilk test ($p > .05$).

4.4 Results

4.4.1 Performance of the Detection Methods

A classification table in the form of a confusion matrix was constructed to evaluate the performance of each artefact detection method. Figure 4.2 demonstrates the confusion matrices for the absolute-fixed, median-fixed, and adaptive threshold techniques against the KUB-detected signal. The absolute-fixed method had the highest misclassification rate of normal instances (2.8%), approximately twice as high as the median-fixed (1.3%) and adaptive threshold (1.5%) approaches.

As outlined in Table 4.1, the adaptive threshold approach obtained the highest performance metrics in terms of balanced accuracy (90.2%) compared to absolute-fixed (81.8%) and median-fixed (84.8%). Similarly, the adaptive threshold had the highest sensitivity score at 81.5%, indicating that 18.5% of the artefacts were misclassified as normal instances. Nonetheless, all detection methods had a similar specificity score, ranging from 97% to 98%.

Lastly, the processing time for each detection method was calculated as the average time taken to identify the artefacts from all data files. The results revealed that the absolute-fixed was the fastest algorithm ($0.091\ \mathrm{ms}$, SD = $0.03\ \mathrm{ms}$),

Figure 4.2: Confusion Matrices for Artefact Detection Methods

(a) Absolute-fixed

Reference \ Prediction	Normal	Abnormal
Normal	18,551 97.2%	537 2.8%
Abnormal	593 33.6%	1,170 66.4%

(b) Median-fixed

Reference \ Prediction	Normal	Abnormal
Normal	18,848 98.7%	240 1.3%
Abnormal	512 29%	1,251 71%

(c) Adaptive Threshold

Reference \ Prediction	Normal	Abnormal
Normal	18,803 98.5%	285 1.5%
Abnormal	326 18.5%	1437 81.5%

Table 4.1: Performance Metric for Detection Methods against the KUB-Detected Signal and Erroneous Signal (%)

Method	Accuracy 95% CI	Sensitivity	Specificity	Balanced Acc.
KUB-Detected Signal				
Absolute-Fixed	94.6 [94.3, 94.9]	66.4	97.2	81.8
Median-Fixed	96.4 [96.1, 96.6]	71.0	98.7	84.8
Adaptive	97.1 [96.8, 97.3]	81.5	98.5	90.2
Erroneous Signal				
Absolute-Fixed	96.8 [96.6, 97.1]	80.4	98.3	89.4
Median-Fixed	98.9 [98.7, 99]	86.5	99.9	93.2
Adaptive	100 [100, 100]	100	99.9	99.9

followed by the median-fixed (0.717 ms, SD = 0.15 ms) and adaptive threshold (14.6 ms, SD = 3 ms) algorithms. The processing time varied significantly among all detection algorithms. Further, there was a particularly notable time difference between both fixed methods ($\leqslant 1$ ms) and the adaptive threshold (14.6 ms) due to the additional computational steps needed to calculate the time-varying threshold (see Equation 4.3). Based on these accuracy and processing time results, a suitable detection method can be employed in future research according to the requirements of the real-time application in question.

For this study, the adaptive threshold was selected as the best artefact detection algorithm among the investigated methods in light of the balanced accuracy performance metric. In the following section, the reliability of the correction methods will be discussed.

4.4.2 Reliability of the Correction Methods

After using the adaptive threshold for artefact detection, the RR recordings were corrected using three different correction methods: deletion, moving window average, and cubic spline interpolation. Subsequently, HRV analysis was performed on all corrected signals, including the KUB-corrected signal. The HRV measures derived from the KUB-corrected signal served as the reference for this analysis. Further, HRV measures for the original and erroneous signals were included in the analysis to better understand the effect of the artefacts. A summary of the average HRV measures obtained by each filtering method is shown in Table C.1, the ICC results are shown in Table C.2, and the Bland-Altman analysis is presented in Figure C.1 (see Appendix C).

Overall, the reliability of all correction methods for MeanRR and SDNN was excellent, with mean absolute error (MAE) of less than .2% and an ICC of 1. Similarly, NN50 and pNN50 demonstrated excellent reliability when corrected with deletion and moving window average (ICC = .99, 95% CI [.98, .99], MAE < 5%). Although the ICC of the two measures resulting from cubic spline interpolation was excellent (ICC > .96), the 95% CI range was wide with a lower CI bound of .70. Further, the MAE rate was greater than 12%, which was high relative to the other correction methods. Additionally, cubic spline interpolation of RMSSD had the highest error rate (MAE = 23.9%) compared to deletion and moving window average (MAE = 16.5%). Finally, the reliability of the RMSSD measure in all correction methods was good (ICC range .79-.82).

The artefact simulation method used in this study was based on the introduction of abrupt and unexpected changes in the HRV signal, resulting in high-frequency components in the spectral domain. As a result, the presence of artefacts had the greatest impact on HF power (ICC = .17, 95% CI [-.02, .15]), while VLF power had the least effect (ICC = .89, 95% CI [.83, .94]), as shown by the erroneous signal results. Nonetheless, all correction methods significantly improved the reliability analysis for all frequency-domain measures: VLF power (ICC range = .96-.97), LF power (ICC range = .86-.91), and HF power (ICC range = .71-.72). Among the investigated correction methods, moving window average had the lowest MAE across all frequency domain measures (see Supplementary Tables C.1 and C.2). In particular, the MAE of the normalised LF and HF power (i.e., LFnu and HFnu) using the moving window average was less than 3.3%, with a good reliability score (ICC = .73) compared to deletion (MAE < 4.9%, ICC = .66) and cubic spline interpolation (MAE < 10.5%, ICC = .78).

Concerning the non-linear methods, all correction methods showed remarkable improvements in the reliability analysis compared to the erroneous signal. On average, deletion (MAE < 5.4%, ICC range = .81-.96) and window average (MAE < 5.4%, ICC range = .74-.96) had comparable results as well as better MAE rates than cubic spline interpolation (MAE < 11.9%, ICC range = .78-.97).

The Bland-Altman analysis was performed to visualise the agreement between the KUB-corrected signal and HRV measures obtained from each correction method. Figure C.1 illustrates the Bland-Altman plots for the five most common HRV measures: MeanRR, RMSSD, SDNN, LF power, and HF power. The Bland-Altman plots depict the mean difference of measurements, represented by a solid line, and the 95% agreement limits, represented by the dashed lines. In general, all correction methods indicated similar agreement levels, with moving window average and cubic spline interpolation showing slightly better agreement than deletion for the frequency-domain measures.

An open-source implementation for the preprocessing algorithms will allow for future integration with wearable sensors so that signals can be cleaned on the spot while flexibly controlling the filtering parameters (e.g., filtering method, threshold level, window size). Accordingly, the next section describes the development of a designated framework to collect HRV data and perform basic filtering techniques in real time.

4.5 Real-Time Heart Rate Variability Framework

To further explore the feasibility of implementing the preprocessing algorithms in real-time HRV data acquisition, an interface was developed to communicate with CorSense over BLE. This interface was implemented as an open-source framework using the Python programming language and Generic Attributes (GATT) services (Bluetooth Architectural Review Board, 2021). The source code for the data acquisition is publicly available in the researcher's repository on Github (Bahameish, 2019). Technically, the interface can be connected to any BLE-based device; however, it was only tested on the CorSense device in this study.

The GATT framework consists of various profiles, with each profile composed of one or more services. Each service has a set of characteristics that indicate its properties and operations. For instance, the Heart Rate Profile enables communication between the GATT server (e.g., HR sensor) and GATT client (e.g., computer, smartphone). Once the connection between the client and server is established, HR measurements are transmitted to the client using 23 bytes of data per BLE packet (Medical Working Group, 2011a; 2011b).

In a BLE packet, the presence of RR intervals was determined based on the format represented in the flag, which is the first byte of the packet (see Appendix B.3). Based on the flag information, the starting index of the RR measurements was then calculated. Lastly, the RR intervals were retrieved by concatenating every two bytes using the little-endian scheme. The results were then converted into milliseconds, as depicted in Algorithm 2. According to the GATT specification supplement, the resolution of the RR-interval is $1/1024^{\text{th}}$ of a second (Bluetooth Architectural Review Board, 2021).

Further, a basic prototype was developed to integrate the CorSense framework with the preprocessing algorithms to identify and correct the artefacts in real time. The median-fixed threshold approach was adopted for artefact detection and moving window average for correction (see Figure 4.3). Both techniques were implemented using a 10-RR interval window preceding the incoming RR interval.

Algorithm 2: Extracting RR packets from CorSense over BLE

```
Data: Bluetooth BLE packets
Input: Data packet and the starting index (i)            # 1st RR-interval
Output: RR intervals
# Ensure client-server connection is established.
function Extract_RR(data, i):
    while i ⩽ length(data) do
        rr ← data[i] + data[i + 1] ≪ 8        # Combine every two bytes
        rr ← rr/1024 * 1000                  # Resolution 1/1024 s
        index ← i + 2                        # update index
    return rr
```

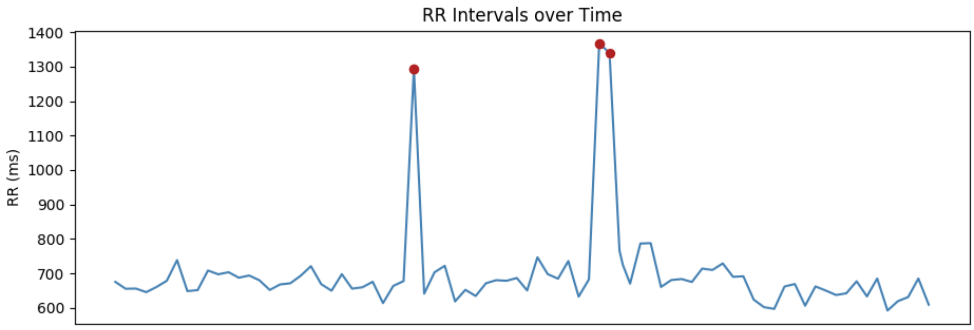


Figure 4.3: A simulation of HRV preprocessing in real-time using CorSense and median-fixed threshold.

4.6 Discussion

There is a consensus among clinicians and researchers regarding the need to remove artefacts from RR intervals for reliable HRV analysis (Peltola, 2012). Although commercial applications provide effective preprocessing algorithms for data filtering, there are limitations in the flexibility of the analysis employed for batch processing or real-time monitoring. Hence, this study examined various filtering approaches in a flexible open-source environment that can provide similar performance to the Kubios application based on two major stages: 1) artefact detection and 2) artefact correction. Moreover, an open-source framework was implemented to integrate real-time HRV data acquisition using BLE-based sensors with the proposed processing algorithms.

Among the investigated artefact detection techniques, adaptive threshold achieved the best performance in artefact identification compared to Kubios (balanced accuracy = 90%, sensitivity = 81.5%) and the erroneous signal (balanced accuracy = 99.9%, sensitivity = 100%). Although the absolute-fixed and median-fixed thresholds had balanced accuracies greater than 81%, both missed approximately one-third of the abnormal instances, with sensitivity rates of 66.4% and 71% respectively. These results are puzzling given that the median-fixed threshold technique was developed to emulate the artefact detection algorithm implemented in the standard Kubios software. However, the performance discrepancy could be attributed to the window size as the local average was calculated based on the median filtering of 10 samples surrounding the artefact. In contrast, the Kubios user guide provides no details regarding the window size of its median filter algorithm (Tarvainen et al., 2016).

For the purposes of developing real-time applications, processing time is as important as detection accuracy. Both fixed-threshold approaches were time-efficient compared to the adaptive threshold approach due to the additional computational steps required to calculate a time-varying threshold for artefact identification. In light of these findings, an appropriate detection algorithm can be selected based on the application's requirements with respect to accuracy and processing time.

While all investigated correction techniques yielded satisfactory HRV measures, the moving window average provided better agreement levels with the Kubios results than deletion and cubic spline interpolation. In addition, the moving window average is computationally efficient and more suitable for real-time applications given the simplicity of its calculations. In contrast, cubic spline interpolation has a higher computational overhead as it requires fitting a higher-degree polynomial and calculating second derivatives (Guven et al., 2016). Deletion resulted in shorter RR intervals, which affected the reliability of the HRV analysis. Comparing these findings with those of other studies confirms that deleting abnormal beats increases the mean error rate for frequency-domain measures (Choi & Shin, 2018; Peltola, 2012; Salo et al., 2001).

Through an open-source implementation, this study aimed to increase the flexibility of the HRV filtering process while maintaining a high level of quality relative to Kubios, a commercially validated application for HRV analysis. The goal was achieved by providing means to control the essential parameters, including filtering method type, threshold value, and window size. Moreover, this study presents approaches to facilitate batch processing for both lengthy HRV recordings and multiple sets of HRV recordings by automating the signal preprocessing methods.

The research to date has evaluated HRV filtering and analysis for real-time purposes by simulating both procedures using an overlapping data segmentation approach with offline processing (Benchekroun et al., 2021; Citi et al., 2012). However, this study extends the previous work by providing a controllable window approach for artefact detection and correction as well as facilitating the integration of BLE-based sensors with real-time HRV data acquisition and filtering. Future practical directions could be explored by fully integrating the two open-source components (i.e., the preprocessing algorithms and CorSense framework) and evaluating the performance of various filtering methods in real-time environments.

4.7 Limitations

In this study, there are three important limitations that must be addressed. First, the simulated artefacts did not include other types of abnormalities, such as ectopic beats resulting from misalignment or extra beats resulting from short RR intervals. Second, the preprocessing performance was evaluated using synthetic artefacts, which left the algorithm's capacity to detect and correct other forms of artefacts unexplored. Third, one correction algorithm was adopted to correct all detected artefacts regardless of abnormality type. A common strategy in offline preprocessing algorithms is to apply the most appropriate correction method based on the identified artefact type (Benchekroun et al., 2022; Citi et al., 2012; Lipponen & Tarvainen, 2019). For instance, if the detected artefact is an extra beat, it is removed; if it is a missed beat, an extra interpolated beat is added. Hence, future research could incorporate this classification scheme when applying the open-source implementation in real-time settings.

4.8 Chapter Summary

This chapter addresses SRQ1 by assessing various artefact detection and correction algorithms to obtain reliable and flexible HRV analysis for future deployment in real-time applications. For the flexibility aspect, this study focused on the provision of an open-source implementation to facilitate flexible adjustment and control of the parameters for each algorithm in accordance with user or application requirements. For the reliability aspect, the results of this study were compared to those obtained from the Kubios application as a benchmark.

The artefact detection algorithms employed in this study presented a trade-off between accuracy and processing time. Ostensibly, advancements in artefact detection accuracy introduced computational overheads (see Section 4.4.1), which then affected processing time. Considering accuracy and processing time, the use of the median-fixed threshold approach for detection and moving window average for correction may provide acceptable results for real-time analysis. However, the adaptive threshold approach is recommended for offline batch processing given its high performance with respect to accuracy.

In an attempt to integrate real-time HRV data acquisition with the preprocessing algorithms, a preliminary open-source interface was developed to communicate with BLE-based HR sensors. This interface employed a window-based approach to identify and correct artefacts in real-time using the median-fixed approach and window average method, respectively.

The next chapter examines the minimum reliable window for HRV analysis to facilitate the development of real-time affective recognition systems, focusing on the batch processing of HRV data collected from several participants at different time points. Hence, the adaptive threshold approach and window average method were employed in the HRV preprocessing stage to ensure high levels of accuracy.

CHAPTER 5

Ultra-Short-Term Analysis

This chapter introduces the significance of ultra-short-term (UST) segments of heart rate variability (HRV) data by outlining the relevant research, followed by an evaluation of the reliability of UST HRV analysis under resting and non-resting conditions (i.e., stress and paced breathing). The influence of stress and paced breathing conditions on HRV is also explored, with special consideration given to measurement consistency across UST segments.

5.1 Overview

UST HRV analysis (< 5 min) has received considerable scholarly attention in recent years given its capacity to provide momentary insights regarding current physiological states. According to standards established by the Task Force (Malik et al., 1996), HRV can be analysed over long periods, with a maximum recording length of 24 h, or short periods, with a minimum recording length of 5 min. While previous research has considered the 5-min duration to be the gold standard for short-term analysis, there is no scientific evidence to support this claim. Quintana et al. (2016b) pointed out that the persistence of this standard is likely due to the processing limitations of the computers used for ECG data acquisition in the 1960s. However, power spectral analysis should be conducted with caution while considering signal characteristics (e.g., sampling frequency, signal stationarity; Li et al., 2019).

Several studies have explored the feasibility of using HRV analysis in periods of less than 5 min to monitor dynamic HR fluctuations. Their primary aim was the incorporation of UST analysis in physical exercise and mental health applications, thereby facilitating future deployment in wearable devices and real-time systems (Castaldo et al., 2015; Lee et al., 2022; Wu et al., 2020). However, most of these studies focused on HRV measures in a resting state.

Using a concurrent validity approach, the present study sought to extend these findings by investigating the reliability of UST HRV segments under resting conditions as well as two additional non-resting conditions (i.e., stress and paced breathing). Concurrent validity is a criterion validity in which reliability is evaluated by comparing new measurements to previously validated standard measurements (Shaffer et al., 2020). In this study, correlation analysis, limits of agreements, and trend analysis were used to compare HRV measures derived from the UST analysis to those obtained from the standard 5-min recording (Malik et al., 1996).

5.2 Related Work

The growing demand for individual mental health improvement combined with the recent development of wearable technologies have brought attention to real-time health monitoring (Can et al., 2020; Hinde et al., 2021). Real-time health monitoring can track and record physical activity, physiological signals, and social interactions. This technology has led to new approaches in the identification of early warning signs for mental health conditions as well as the facilitation of timely clinical interventions, thus preventing conditions from worsening. For instance, leveraging the sensing abilities of wearable devices during physical activity can provide indicative insights regarding the symptoms of bipolar disorder and depression (Abdullah & Choudhury, 2018).

As discussed earlier, HRV provides a good indication of stress levels and physical health; however, a minimum window of 5 min does not conform to real-time requirements. Accordingly, researchers have investigated the reliability of UST analysis during resting states. For instance, Sheridan et al. (2021) used the Pearson correlation coefficient to assess correlations between commonly assessed HRV measures (e.g., HF power, LF power, RMSSD, SDNN) using a 5-min HRV recording and the first minute of that recording. They found a high correlation between HF power and RMSSD ($r > .90$ for both), whereas moderate correlation coefficients were shown in SDNN ($r = .63$) and LF power ($r = .71$). However, the study had a number of drawbacks, including a small sample size of four participants, details about participants' conditions during the assessed segments not being reported, and sole reliance on Pearson correlation analysis.

In a review of research on the reliability and validity of HRV measures resulting from UST analysis, Pecchia et al. (2018) demonstrated that correlation analysis is not sufficient for concluding that an HRV measure derived from the UST analysis is a reliable estimate for the 5-min standard. In light of the identified methodological ambiguities and inconsistencies, the authors proposed a set of protocols for appropriate and rigorous assessment of UST segment reliability compared to a benchmark: 1) correlation to examine the association, 2) Bland-Altman analysis to visually inspect bias between means, 3) statistical significance among resting and stress conditions, and 4) trend analysis to ensure consistency across time segments during stress conditions.

In seeking to develop an auto-detect stress classifier, Castaldo et al. (2019) followed the reported guidelines and discussed the results of time-domain, frequency-domain, and non-linear measures in 3-min, 2-min, 1-min, and 30-s periods during resting and stress phases. They found that MeanRR, SDNN, HF, and SD2 presented great consistency at the 1-min segment as well as a high correlation with the 5-min recordings. In the 30-s segments, the computation of some HRV measures (e.g., LF, HF, LF/HF) led to erroneous values due to the insufficient number of samples. In fact, Malik et al. (1996) pointed out that the length of the segment in the spectral analysis should be 10 times the wavelength of the lower bound frequency of the investigated spectral band in spectral analysis. For example, to obtain a reliable analysis of HF power (frequency band .15-.40 Hz), the minimum length should be around 1 min, which is calculated as $10 \times \frac{1}{.15} \approx .66$ seconds. Nevertheless, Shiraishi et al. (2018) provided a promising visualisation approach for the power spectrum of HRV in real-time during exercise. They selected a moving window of 30 s updated with every heartbeat, and the analysis was performed using maximum entropy rather than the FFT or autoregressive methods.

Munoz et al. (2015), a seminal work in this field, provided a thorough reliability test for UST analysis of segments with durations of 2 min, 30 s, and 10 s using Bland-Altman analysis, Cohen's d, and correlation analysis. At three averaged 10-s segments, they achieved high correlation coefficients with the standard 5-min intervals (RMSSD: r = .94; SDNN: r = .86). However, this study was limited to the RMSSD and SDNN measures taken from the time domain during a resting condition. In contrast, Shaffer et al. (2016) examined UST reliability in a resting condition by adopting strict validity criteria for Pearson's r (r > .90). They found that a 1-min segment was required for reliable estimation of time-domain measures (e.g., RMSSD, SDNN, NN50, pNN50), a 2-min segment for LF power, and a 3-min segment for HF power and LF/HF.

A recent study by Burma et al. (2021) investigated the reliability of UST HRV measures by extracting 240-, 180-, 120-, 60-, and 30-s segments from a 5-min recording during a resting condition. They employed concurrent validity using Bland-Altman plots with 95% limits of agreement, coefficient of determination, coefficient of variation, and repeated measures ANOVA. They found that RMSSD and SDNN in segments of less than 240 s showed low levels of agreements with the 5-min window. Hence, they concluded that a minimum of 4 min (240 s) is the shortest recommended window for reliable UST HRV analysis. However, the scope of the study was primarily concerned with an upright orthostatic position. Therefore, the generalisability of these findings on supine or seated positions is limited due to the impact of posture on the SNS (Quintana et al., 2016b).

Additionally, Melo et al. (2018) evaluated the reliability of 180-, 120-, and 60-s UST segments against a 5-min RR interval under two breathing conditions using the Pearson's correlation and statistical group mean differences test. The authors found that RMSSD in paced breathing had a higher correlation coefficient compared to spontaneous breathing, and this correlation was reliable at the 60-s segment. Although a t-test analysis revealed a greater difference in the means of paced breathing compared to spontaneous breathing, no considerations were given to the limits of agreements nor quantification of an acceptable difference between measurements. Moreover, the shortest duration investigated was 60 s.

Expounding upon the comprehensive review conducted by Pecchia et al. (2018), Shaffer et al. (2020) critically addressed the limitations of various analytical comparison approaches utilised in 28 studies. They argued that Bland-Altman analysis with a priori is the most appropriate analysis to identify acceptable agreement levels between the 5-min RR interval and shorter segments. Accordingly, the reliability of UST analysis is investigated in the present study under resting, stress, and paced breathing conditions using Bland-Altman analysis with a priori, Pearson's correlation, and trend analysis. The latter was used for non-resting conditions to ensure trend consistency. Moreover, the influence of mental stress and paced breathing conditions was analysed for the 5-min segment and across UST segments. Accordingly, this study sought to address the following research question:

> **SRQ2:** What are the requirements for a reliable real-time HRV analysis using UST segments under resting, stress, and paced breathing conditions?

5.3 Hypotheses

The hypotheses associated with SRQ2 are as follows:

> **Hypothesis 1 (H1):** UST analysis can provide reliable HRV measures compared to the 5-min standard under resting, stress, and paced breathing conditions.

> **Hypothesis 2 (H2):** HRV measures under non-resting conditions (i.e., stress and paced breathing) differ from those under resting conditions, and these HRV measures are consistent across UST segments in each condition.

The reliability of H1 was defined as the degree to which HRV measures resulting from UST segments could accurately and consistently estimate the corresponding measures of the 5-min standard recording using concurrent validity.

5.4 Methods

5.4.1 Participants

A sample size of at least 20 was deemed necessary for 80% statistical power in detecting a medium effect size of .3, a significance level of .05, and a correlation among repeated measures of .50, as calculated using a priori power analysis in G*Power for a one-way repeated measures ANOVA (Erdfelder et al., 1996). While determining the target sample size, the number of participants involved in similar studies was also considered (i.e., a minimum of 20 participants; Esco & Flatt, 2014; Melo et al., 2018; Salahuddin et al., 2007a).

Twenty-four healthy participants (aged 20-36 years) from QMUL were recruited to participate in the study based on a call for participation sent to students and faculty in the computer science department. To minimise external factors that could affect cardiac activity measurements, participants were instructed to avoid caffeine, smoking, and eating heavy meals for 2 h prior to the study as well as engaging in intense physical workouts for 24 h prior to the experiment (Laborde et al., 2017; Quintana et al., 2016b). As recommended by Quintana et al. (2016b), participants completed a simple questionnaire asking about their alcohol/coffee intake, fitness level, sleep routine, and overall physical health (see Appendix A.2). In addition, the GAD-7 questionnaire was given to all participants to ascertain the impact of self-reported anxiety on HRV levels (see Section 3.3.2 and Appendix A.3; Quintana & Heathers, 2014). The study was approved by the Queen Mary Research Ethics Committee (QMERC2019/58). All participants were informed about the nature of the experiment and signed a written consent form. The documents related to ethical approval and consent are shown in Appendix D (see Figures D.1 and D.2).

After the data filtering process, four participants were excluded due to a high percentage of artefacts in the signal exceeding 5% of the RR recording. Hence, the analysis was conducted with 20 participants: 11 men (age: M = 27.6 $\pm$ 4.3 years) and nine women (age: M = 27.9 $\pm$ 2.2 years).

5.4.2 Experiment Design

The study was based on a repeated measures quasi-experimental design aimed at collecting HRV data (dependent variable) in three conditions (independent variable): 1) before the study as a **baseline** measurement, 2) during mental **stress** task, and 3) during **paced breathing**. The paced breathing was performed by all participants as a post-stress recovery exercise to improve HRV and reduce stress. HRV was measured using the CorSense device, a PPG-based sensor with a sampling rate of 500 Hz (see Section 3.3.3).

Prior to the study, additional demographic and affective information were collected using the GAD-7 questionnaire. Further, participants were asked to rate their experience with deep breathing, fitness level, and physical activity using a five-point single-item Likert scale. To facilitate the comparison of HRV changes, all participants experienced the same conditions in the same order from baseline to paced breathing after going through a stress phase. This order is in keeping with the 3 Rs approach (i.e., resting, reactivity, and recovery) proposed by Laborde et al. (2017).

Trier Social Stress Task

Widely implemented in laboratory settings, the TSST protocol includes speaking and arithmetic tasks that can effectively induce stress in individuals (Kirschbaum et al., 1993). This study adapted the TSST protocol to elicit a stress response in participants. First, the speaking task had 10 different topics, such as education, environment, family, travel, and work. Participants had to choose a random number from one to 10 to select a topic. Each topic included three questions written on small cards to facilitate speaking (e.g., “Do you think job satisfaction is more important than salary when choosing a job?”, “What skills do you think are needed to get a good job these days?”, “What jobs do you think are most valuable to society?”). Second, the arithmetic task involved a mental sequential subtraction activity with standardised initiation and subtraction numbers. Prior to the study participants were generally informed that the HRV data would be collected during three conditions: resting, mental task, and paced breathing. The details of the TSST protocol were provided during the phase of the mental task.

Heart Rate Variability Processing

The collected HRV data passed through four phases of signal processing: 1) filtering, 2) quality check, 3) segmentation, and 4) HRV analysis. A schematic diagram illustrating these phases is shown in Figure 5.1.

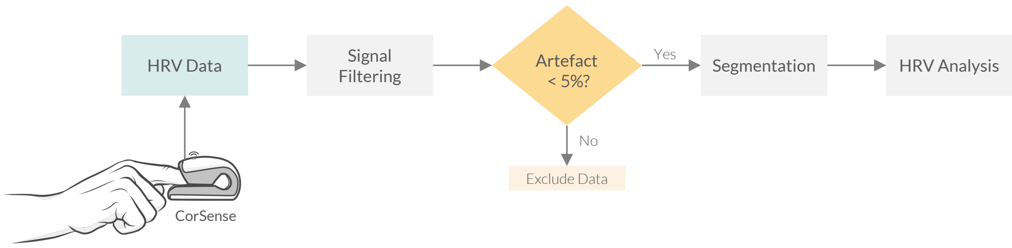


Figure 5.1: HRV Signal Processing Phases

Initially, an exploratory HRV analysis was conducted using the Kubios application. However, Kubios does not support batch processing and controlled window analysis. Thus, signals were filtered using the adaptive threshold artefact detection and moving window average correction methods, according to the implementation discussed in Chapter 4 (see Section 4.8). Subsequently, the HRV analysis for the time-domain and frequency-domain analyses was performed in Python using Systole (Legrand & Allen, 2022), while the HRV analysis for non-linear analyses was performed using pyHRV (Gomes, 2018). Systole provides an exclusive implementation of the Welch's periodogram, an FFT-based method, for power spectral estimation. Thus, the frequency-domain measures in this study were computed using Welch's method.

For the stress condition, HRV was separately recorded for the speaking and arithmetic tasks; however, these were later combined into one signal as the paired-sample t-test showed non-significant differences in HRV measures between both stress tasks ($p > .05$ for all). Subsequently, a 5-min interval was extracted from the centre of the combined HRV signal during analysis.

Data Analysis Approach

For this experimental study, the hypothesis regarding the reliability of UST analysis (H1) was assessed based on concurrent validity with 5-min RR-intervals using Bland-Altman analysis with 95% limits of agreement, Pearson's correlation coefficient, and trend analysis. The latter was used to ensure the consistency of changes from baseline to non-resting conditions. To assess the hypothesis regarding significant changes in HRV measures based on varying experimental conditions (H2), MLL model analysis was applied, followed by Tukey's post-hoc analysis of pairwise comparisons. An overview of the statistical analysis approach used for both hypotheses is depicted in Figure 5.2.

As a measure of effect size, the omega squared (ω^2) was reported for the MLL analysis. The values of .01, .06, and .14 were interpreted as small, medium, and large effect sizes, respectively (Cohen, 1988). For the dependent pairwise comparisons, the correlation coefficient (r) was reported with the values of .1, .3, and .5, indicating small, medium, and large effect sizes, respectively (Cohen, 1988).

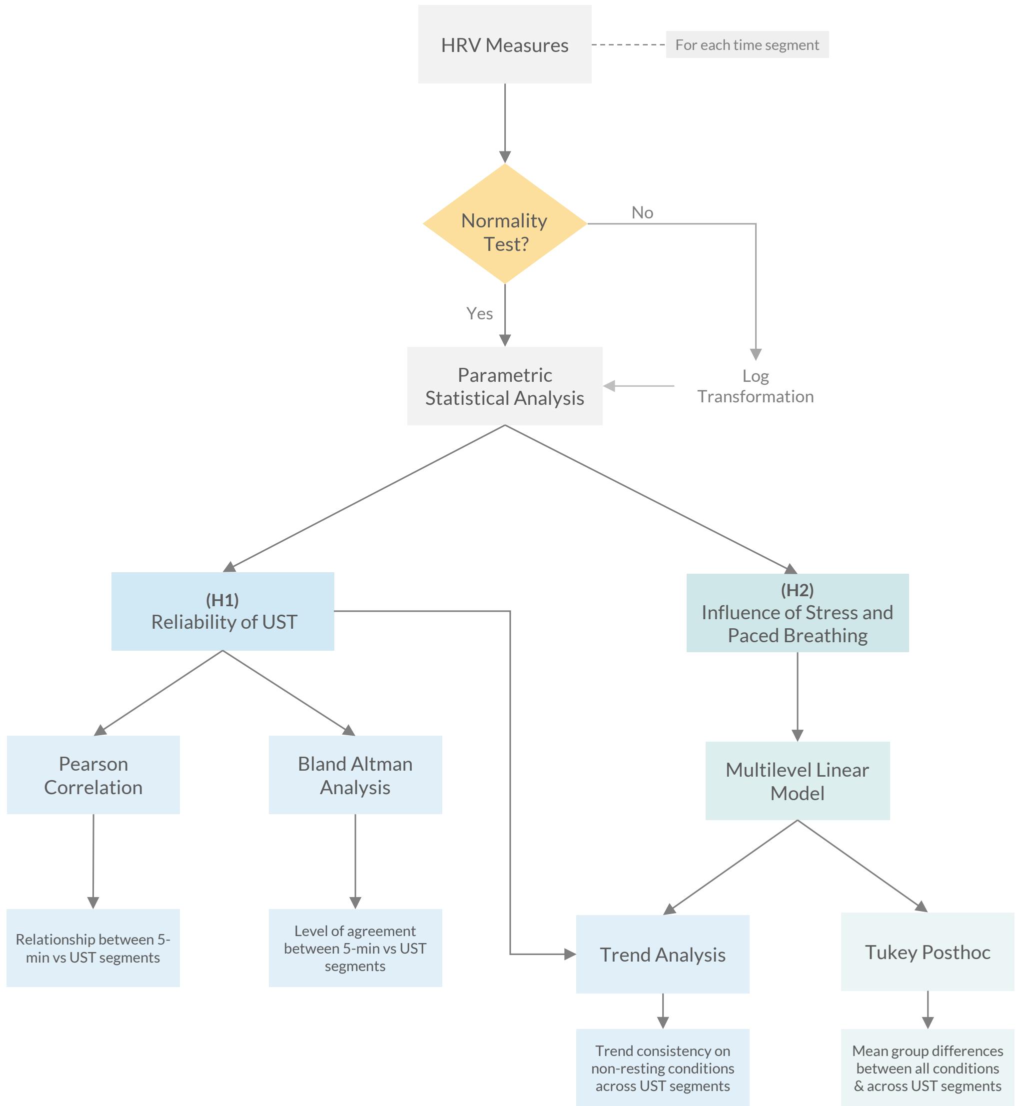


Figure 5.2: A Flowchart for the Statistical Analyses Approaches

5.4.3 Procedure

The experiment sessions were conducted in the daytime (9:30 am to 12:30 pm) over several days, and each session lasted 35 min. The experimental study consisted of HRV collected under 1) a controlled condition to establish a baseline measurement during which participants were instructed to sit quietly and breathe naturally for 6 min, 2) a mental stress task based on TSST with a duration of 15 min, and 3) a paced breathing exercise for 6 min as a post-stress activity.

During the stress task, participants had 5 min to mentally prepare for the speaking task without written notes, and then 5 min to deliver a speech. The last 5 min were dedicated to the mental arithmetic task in which participants were asked to subtract 13 from 1,022 sequentially. All TSST protocols were followed while interacting with the participants, as recommended by Birkett (2011). For instance, if a participant paused for 20 s during the speaking task, they were prompted to continue speaking with the phrase: "You still have time remaining". In addition, if a mistake was made in the arithmetic task, the participant was asked to start over from the beginning.

The post-stress task activity was a paced breathing exercise lasting 6 min, which was designed to reduce stress via HR regulation and increase HRV. To breathe at a rate of 7 breaths/min with an equal inhalation-to-exhalation ratio (Lin et al., 2014), participants were asked to follow an illustrative opening and closing circle guide from the Elite HRV application. Each HRV measurement was preceded by a 20-s stabilisation period to allow HR to level out. HRV was measured for 6 min to ensure a minimum recording length of 5 min.

5.5 Data Analysis

5.5.1 Exploratory Analysis

A snapshot of the RR interval and spectral density from one participant during baseline, stress, and paced breathing conditions are shown in Figure 5.3. Closer inspection of this figure shows that the average HRV was slightly lower in stress compared to baseline and paced breathing. Moreover, the signal during paced breathing followed a regular rhythmic pattern, indicated by a peak in the LF band at around 0.1 Hz.

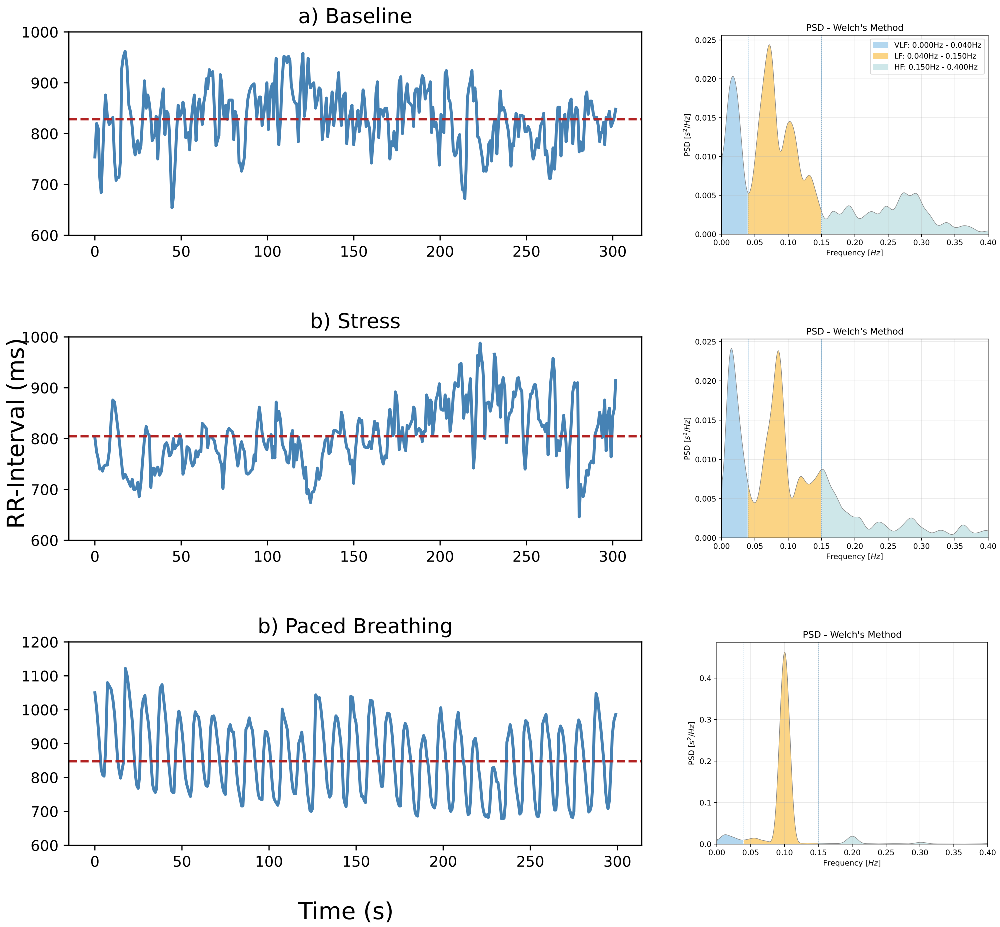


Figure 5.3: A sample of 5-min recording of the RR-Intervals under a) baseline, b) stress, and c) paced breathing

In addition, a time-frequency analysis was performed to visualise changes in the power spectrum with respect to time. Figure 5.4 demonstrates an example of the time-frequency analysis performed on data drawn from one participant during baseline, stress, and paced breathing. During baseline and stress, the frequency components were distributed between the HF and LF bands; conversely, the frequency components were localised in the upper frequency of the LF band (i.e., around 0.15 Hz) during the full recording period under the paced breathing condition. This analytical approach can assess how well participants followed the breathing guide, thus leading to further improvements in breathing regulation and consistency.

A preliminary analysis using histogram plots and the Shapiro-Wilk test showed that HRV measures were not normally distributed ($p < .05$); thus, data were logarithmically transformed. For simplicity and ease of interpretation, HRV data are presented in their non-transformed form (i.e., absolute values) in this chapter, while the log-transformed data are included in Appendix D (see Tables D.1-D.6).

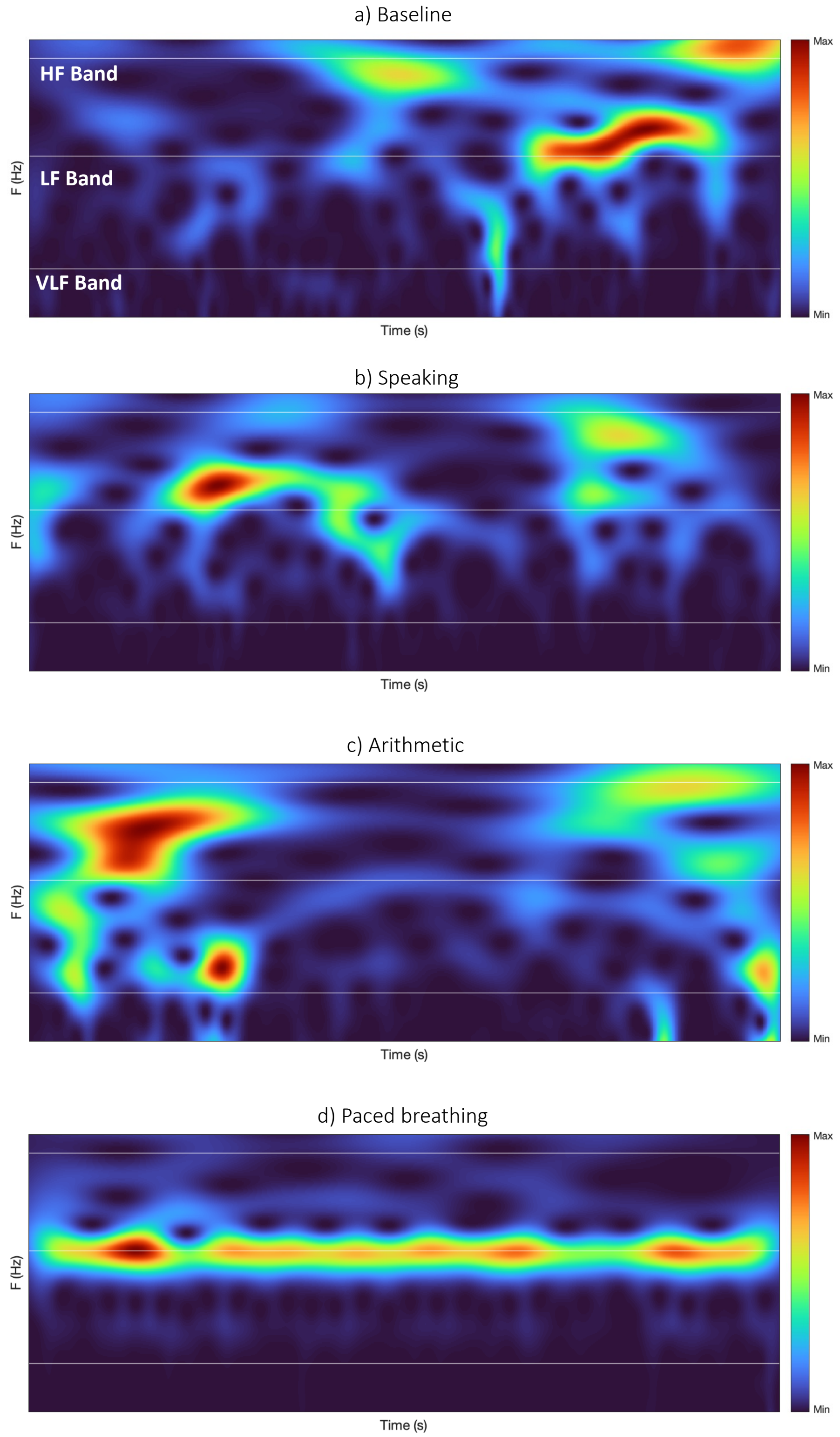


Figure 5.4: Time-Frequency Analysis for a) Baseline, b) Speaking, c) Arithmetic and d) Paced Breathing.

Note. Created with HRVAS toolkit in Matlab (Ramshur, 2010)

5.5.2 Descriptive Statistics

A descriptive statistical overview of participants' demographic characteristics is provided in Table 5.1, while a summary of the average HRV measures over the 5-min interval for each condition (baseline, stress, and paced breathing) is shown in Table 5.2.

Table 5.1: Demographic Information Summary (N=20)

Age	GAD-7	Deep Breathing	Physical Activity	Fitness Level	Sleep Hours
27.9 ± 5.7	3.5 ± 4.3	2.3 ± 1.2	3.35 ± .8	3.15 ± .8	7.2 ± 1.1

As assessed by the Pearson's correlation coefficients, there were no significant relationships between the GAD-7 score and HRV measures in all conditions or the remaining survey questions ($p > .05$ for all). Although there were no significant mean differences related to gender among all baseline HRV measures ($p > .13$ for all), the average baseline RMSSD, and HF power were higher in women than men, as shown in Figure 5.5 (see Section 2.2.6).

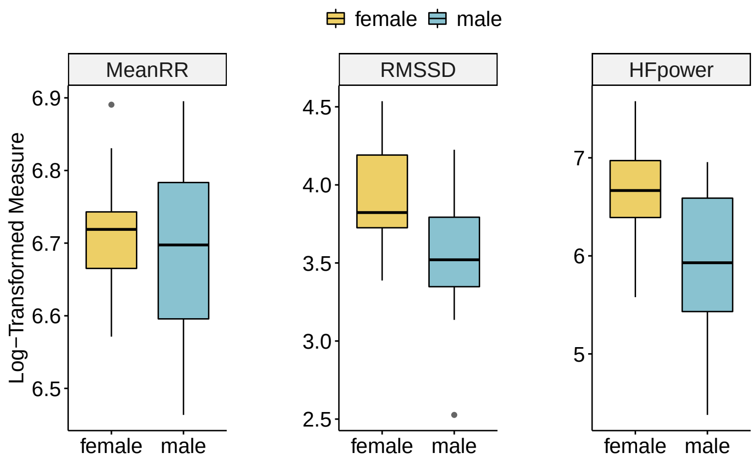


Figure 5.5: Boxplots for the Baseline MeanRR, RMSSD, and HF Power based on Gender.

Table 5.2: Summary Statistics of HRV Measures in the 5-min Recording

Measure	Baseline		Stress		Breathing	
	M	SD	M	SD	M	SD
Time-Domain						
MeanRR	820.8	94.0	737.8	103.3	830	119.7
RMSSD	47.3	19.2	44.1	32.2	57	27.5
SDNN	70.8	24.4	78.3	45.5	100	37.5
pNN50	18.8	12.5	18.5	13.1	28	16.4
Frequency-Domain						
VLF power	1316.7	1045.8	1529.8	1777.8	4289.8	13354.0
LF power	1859.2	1219.8	2487.7	2279.0	6762	5415.3
HF power	681.3	469.4	1277.3	1937.1	1003	843.8
LFnu	72.2	12.7	72.6	13.3	85.7	11.3
HFnu	27.8	12.7	27.4	13.3	14.3	11.3
LF/HF	3.4	2.3	3.8	2.9	9.0	5.5
Total power	3857.3	2264.9	5294.8	5727.3	12054.8	15444.2
Non-Linear						
SD1	31.2	13.6	35.2	22.8	43.8	19.4
SD2	94.8	32.0	104.4	60.7	133.0	50.2
SampEn	1.4	0.2	1.2	0.3	1.0	0.3
DFA1	1.3	0.1	1.3	0.1	1.4	0.2
DFA2	0.8	0.2	0.7	0.1	0.6	0.2

5.5.3 Reliability of Ultra-Short-Term Analysis

Method

Each HRV recording in the dataset was divided into shorter time segments of 120, 60, 30, 20, and 10 s. The segmentation process was conducted by randomly selecting a starting point within the recording and then proceeding in a recursive manner. Figure 5.6 illustrates the segmentation process, where:

- st_0 is the starting time of the 5-min recording,
- st_i is the starting time of the subsequent segment,
- w_i is the length of the window segment, and
- end_i is the ending time point.

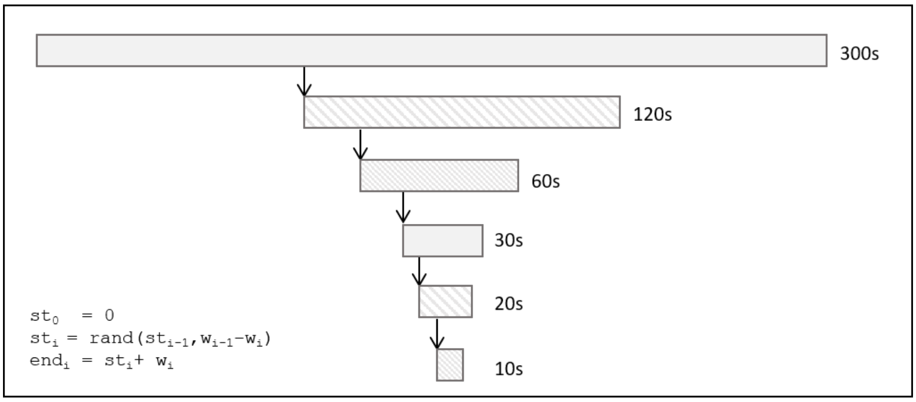


Figure 5.6: HRV Data Segmentation Process

At each UST segment, the ICC was used as a test-retest reliability assessment for the HRV measures obtained from the random segmentation of the RR intervals (Koo & Li, 2016). There was excellent agreement among the five generated samples, with an ICC of .97 and a 95% CI between .969 and .971 ($p <$ .001). Accordingly, one of these randomly generated samples was selected for the analysis in the present study. According to Table 2.1, the HRV measures were compared with the 5-min standard as a benchmark using Bland-Altman analysis, Pearson's correlation coefficient, and trend analysis.

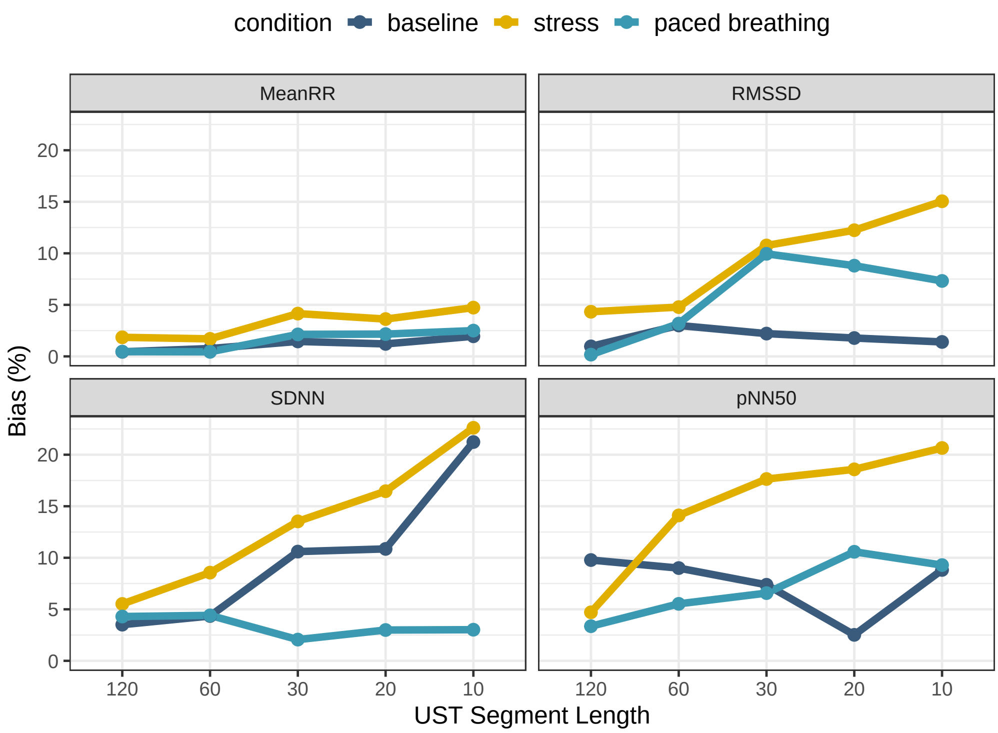
Long-term HRV recordings are recommended to compute ULF and VLF power for a reliable analysis (Malik et al., 1996); thus, these two measures were excluded from this study. In addition, NN50 was excluded because it depends on the absolute number (i.e., count) for the difference between adjacent RR intervals that exceed 50 ms. Moreover, SD1 and SD2 from the non-linear methods were included for the 120- and 60-second segments because the measures in ultra-shorter segments could not otherwise be computed (Chou et al., 2021).

For Bland-Altman analysis, the Shapiro-Wilk test demonstrated that the differences in measurements were normally distributed ($p > .05$). The Bland-Altman bias, or acceptable mean difference, was set to 10% as a priori. In the context of this study, the strength of the correlation was interpreted as follows: 0-.19 negligible, .20-.39 very weak, .40-.59 weak, .60-.79 moderate, .80-.89 strong, and .90-1 very strong (Schober & Schwarte, 2018).

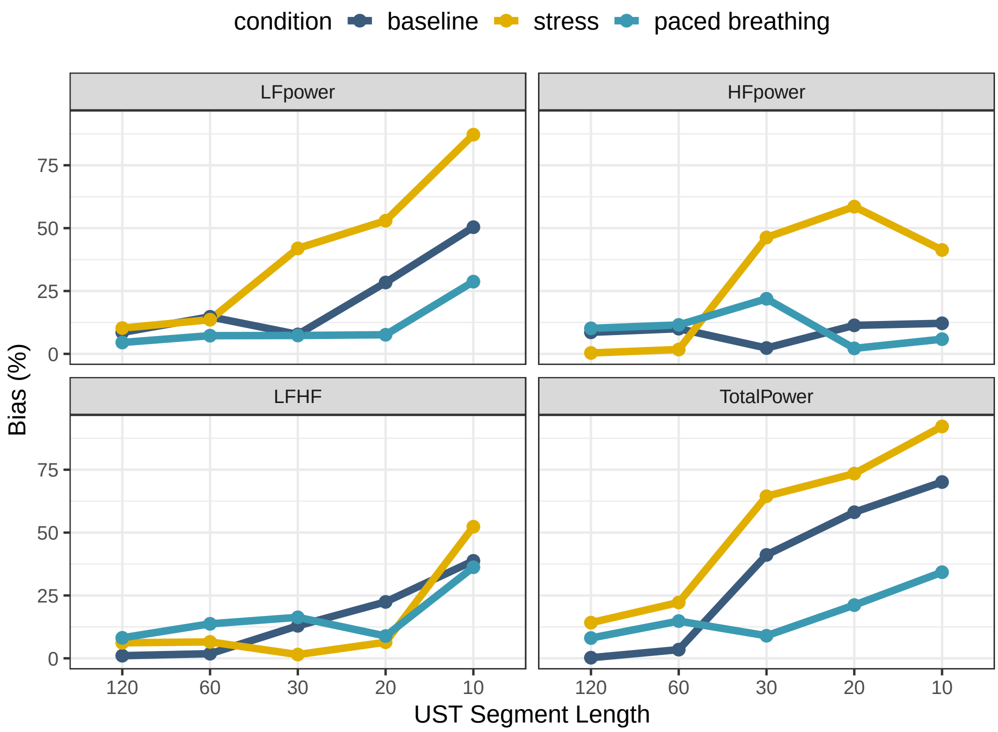
Results | Bland-Altman Analysis

Bland-Altman analysis was performed across all HRV measures and conditions as a visual approach to assess agreement levels. Supplementary Figures D.3-D.8 show the results of the analysis with 95% limits of agreement.

Moreover, bias in the Bland-Altman analysis (i.e., mean difference) with a 95% CI was calculated as a percentage to compare the results with a pre-determined a priori of 10% (see Supplementary Tables D.7-D.9). Figure 5.7 demonstrates bias trends across the UST segments in all conditions. Generally, the bias of all HRV measures noticeably increased as the segments became shorter.



(a) Time-Domain Measures



(b) Frequency-Domain Measures

Figure 5.7: Bland-Altman bias (%) of HRV Measures in Baseline, Stress, and Paced Breathing.

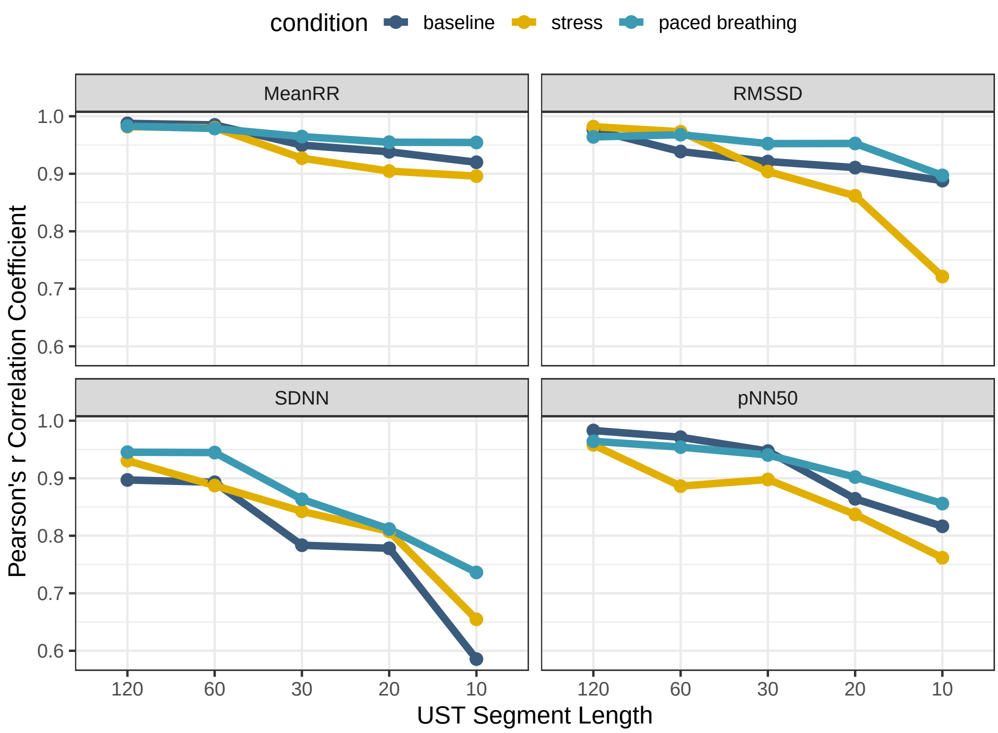
Although the bias of MeanRR followed the general trend of increasing bias with decreasing length of the segment, the bias of all UST segments was less than 5% across all conditions. During baseline, RMSSD maintained a bias of less than 5% across all segments. However, RMSSD bias increased in segments of less than 60 s during stress and paced breathing (30 s: bias > 10%). The bias of SDNN increased at the 30-s segment during baseline and stress, with a bias greater than 10%. Conversely, SDNN maintained a bias of less than 5% across all segments during paced breathing. The pNN50 maintained a bias of less than 10% during baseline and paced breathing across all segments, while pNN50 bias was significantly high at the 60-s segment in stress condition (14.1%).

The bias of the frequency-domain measures was higher than that of the time-domain measures. The bias of HF power at the 60-s segment during baseline was 10%. LF power maintained a bias of less than 10% up to the 20-s segment during paced breathing, while LF bias was significantly high at the 20-s segment in resting and stress conditions (28% and 52%, respectively). Total power had a similar trend to LF power, with bias increasing significantly as the segments became shorter; further, bias was lower in paced breathing compared to the other conditions. Lastly, SD1 and SD2 from the non-linear methods showed a bias of less than 10% at the examined segments (i.e., 120 s and 60 s).

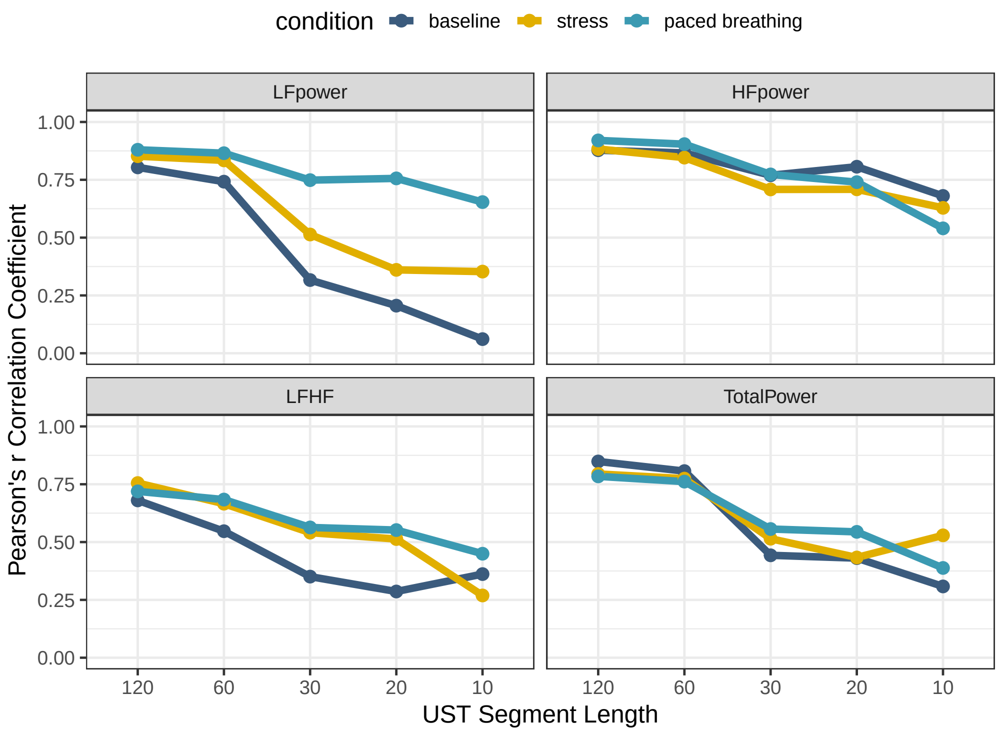
Results | Pearson's Correlation Analysis

For each experimental condition, Pearson's correlation was performed on the log-transformed HRV measures obtained from the UST segments and the 5-min RR segment (see Figure 5.8 and Supplementary Tables D.10-D.12).

Out of all the HRV measures, MeanRR maintained very strong relationships with the 5-min interval across all UST segments in all conditions ($r > .90$ for all, $p < .001$ for all). The remaining HRV measures showed a similar trend, with



(a) Time-Domain Measures



(b) Frequency-Domain Measures

Figure 5.8: Pearson's correlation coefficient of HRV measures in Baseline, Stress, and Paced Breathing.

the correlation coefficient decreasing as the segment became shorter. During baseline, RMSSD and SDNN showed strong relationships at the 120-s segment (r_{RMSSD} = .98, 95% CI [.97, 1.0]; r_{SDNN} = .90, 95% CI [.82, .96]). However, the correlation coefficient decreased at the 10-s segment, with RMSSD and SDNN showing strong and weak relationships, respectively (r_{RMSSD} = .89, 95% CI [.73, .96]; r_{SDNN} = .59, 95% CI [0.19, 0.82]). The pNN50 measure presented a similar pattern to the RMSSD.

Concerning the frequency-domain measures, the segments shorter than 60 s showed a substantial drop in correlation coefficients as segment length decreased. For instance, total power exhibited strong (r = .81) and weak relationships (r = .44) at the 60-s and 30-s segments during baseline, respectively. In addition, LF/HF showed correlation coefficients of less than .76 across all time segments in all conditions. Conversely, HF power maintained strong correlation coefficients up to the 60-s segment in all conditions. Regarding the non-linear measures, SD1 and SD2 had strong relationships at the 60-s segment across all conditions ($r > .86$ for all, $p < .001$ for all).

Further, the correlation varied based on the condition in which the data were collected. For instance, SDNN and LF power during paced breathing showed higher correlation coefficients at the 10-s segment ($r_{SDNN} = .74$, 95% CI [.48, .85]; r_{LF} =.65, 95% CI [.40, .85]) compared to resting and stress conditions (r_{SDNN} < .65, r_{LF} < .35). Moreover, the correlation of RMSSD during stress was slightly lower than that obtained at segments below 30 s in the other conditions, as demonstrated in Figure 5.7.

5.5.4 Influence of Stress and Paced Breathing on Heart Rate Variability

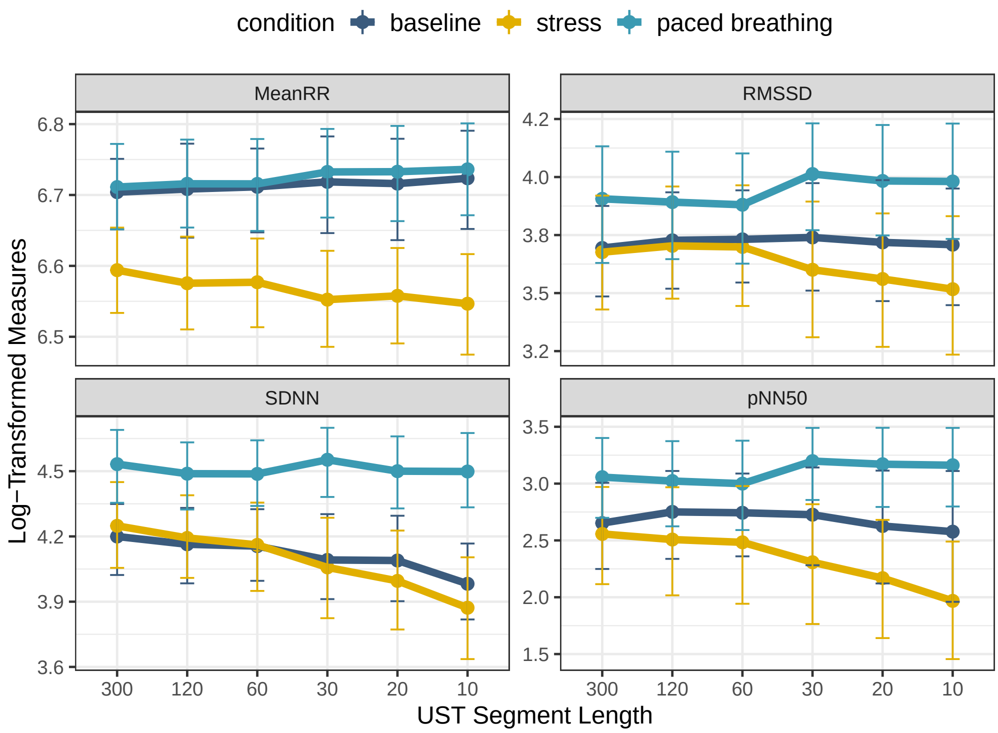
Method

An MLL model was used to analyse the impact of the conditions (resting, stress, and paced breathing) on the HRV measures. This model was selected over one-way repeated measures ANOVA due to violation of the sphericity assumption in which the variances of the differences among all pairs were unequal, as assessed with Mauchly's test (W = .62, $p < .05$, $\epsilon = .72$). Furthermore, Tukey's post-hoc analysis was conducted to perform pairwise comparisons of the HRV measures between the different conditions. The significance level was established as 5%.

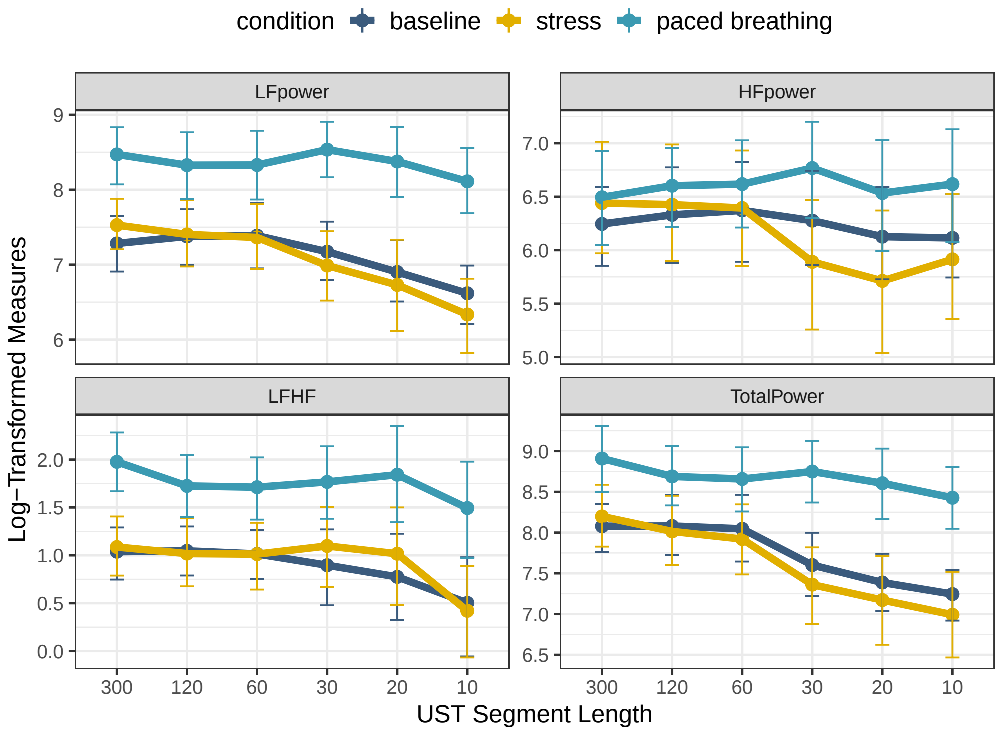
Results

The mean of HRV measures based on the experimental condition is presented in Figure 5.9. In addition, the results of the MLL analysis for all HRV measures across all time segments are summarised in Supplementary Table D.13. In this section, the statistical results discussed are those of the 5-min segment, unless otherwise stated.

The experimental condition had a significant effect on HRV measures across all time segments, including MeanRR and SDNN from the time-domain method (MeanRR $\chi^2(2) = 54,\ p < .001,\ \omega^2 = .72$; SDNN $\chi^2(2) = 12,\ p = .003,\ \omega^2 = .22$). However, there were no significant changes in RMSSD ($\chi^2(2) = 4.3,\ p = .12,\ \omega^2 = .05$). Although there was a significant effect on pNN50 at 5 min, it was not consistent across all UST segments (5 min: $\chi^2(2) = 7.6,\ p = .02,\ \omega^2 = .13$; 120 s: $\chi^2(2) = 5.2,\ p = .07$). Similarly, there were significant effects on the frequency-domain measures across all time segments (LF: $\chi^2(2) = 25,\ \omega^2 = .43$; LF/HF: $\chi^2(2) = 22,\ \omega^2 = .39$; total power: $\chi^2(2) = 13,\ \omega^2 = .24$; $p < .001$ for all). Moreover, SD1 and SD2 were statistically significant between the conditions across the investigated time segments ($p < .05$ for all). By contrast, the results



(a) Time-Domain Measures



(b) Frequency-Domain Measures

Figure 5.9: Mean of HRV Measures with 95% CI in Baseline, Stress, and Paced Breathing across all UST Segments

of HF power were not significant across all time segments, except for the 30-s segment (5 min: $\chi^2(2) = 1.1$, $p = .6$, $\omega^2 = -.03$; 30 sec: $\chi^2(2) = 8.2$, $p = .02$).

According to the Tukey's post-hoc results, MeanRR was significantly lower during the mental stress task at the 5-min segment compared to baseline ($b = -0.11$, $t(38) = -8.8$, $p < .001$, $r = .82$) and paced breathing ($b = -0.12$, $t(38) = -9.3$, $p < .001$, $r = .83$), representing large effect sizes. However, there were no significant changes in MeanRR between baseline and paced breathing ($t(38) = .6$, $p = .84$, $r = .10$). In contrast, SDNN increased significantly during paced breathing compared to baseline ($b = .33$, $t(38) = 3.4$, $p < .001$, $r = .48$) and compared to stress ($b = .28$, $t(38) = 2.9$, $p = .02$, $r = .43$). Similarly, pNN50 increased from stress to paced breathing ($b = .50$, $t(38) = 2.7$, $p = .03$, $r = .40$). Although there was no significant change in RMSSD, the average RMSSD in paced breathing was relatively higher than baseline ($b = .21$, $t(38) = 1.7$, $p = .21$, $r = .27$) and stress ($b = .23$, $t(38) = 1.9$, $p = .16$, $r = .29$), representing medium sized-effects, respectively.

Further, a significant increase in LF power was found during paced breathing compared to baseline ($b = 1.19$, $t(38) = 5.4$, $p < .001$, $r = .66$) and paced breathing compared to stress ($b = .94$, $t(38) = 4.3$, $p < .001$, $r = .57$). Similarly, LF/HF and total power showed a significant increase during paced breathing compared to baseline and stress ($p < .05$ for all). Although there was no significant change in HF power, the average HF power in paced breathing was relatively higher than baseline ($b = .25$, $t(38) = .90$, $p = .62$, $r = .15$), representing a small effect size. Regarding the non-linear methods, a significant increase was shown in paced breathing compared to baseline in SD1 ($b = .32$, $t(38) = 2.8$, $p = .02$, $r = .41$). Moreover, SD2 was significantly higher in paced breathing compared to baseline ($b = .33$, $t(38) = 3.3$, $p = .01$, $r = .47$) and paced breathing compared to stress ($b = .29$, $t(38) = 2.9$, $p = .02$, $r = .43$).

5.5.5 Heart Rate Variability Measures Trend Analysis

Method

Following the MLL analysis, trend consistency was explored across all time segments. The change in HRV means from baseline to stress and from baseline to paced breathing determined the direction of change, which is represented by the t-statistic score retrieved from the post-hoc analysis. The p-value was used to indicate the significance of the change.

Results

Table 5.3: Trend Analysis

Measure	Stress						Breathing					
	300	120	60	30	20	10	300	120	60	30	20	10
MeanRR	⇊	⇊	⇊	⇊	⇊	⇊	↑	↑	↑	↑	↑	↑
RMSSD	↓	↓	↓	↓	↓	↓	↑	↑	↑	↑	↑	↑
SDNN	↑	↑	↑	↓	↓	↓	⇈	⇈	⇈	⇈	⇈	⇈
pNN50	↓	↓	↓	↓	↓	↓	⇈	↑	↑	⇈	↑	⇈
LF power	↑	↑	↓	↓	↓	↓	⇈	⇈	⇈	⇈	⇈	⇈
HF power	↑	↑	↑	↓	↓	↓	↑	↑	↑	↑	↑	↑
LF/HF	↑	↓	≈	↑	↑	↓	⇈	⇈	⇈	⇈	⇈	⇈
Total power	↑	↓	↓	↓	↓	↓	⇈	⇈	⇈	⇈	⇈	⇈
SD1	↑	↓	↓				⇈	↑	↑			
SD2	↑	↑	↑				⇈	⇈	⇈			

The trends in the average HRV measures during stress tasks and paced breathing were investigated for each UST segment. Table 5.3 summarises the results, with arrows (↑, ↓) indicating the direction of change from baseline (i.e., increase or decrease), double arrows (⇈, ⇊) indicating the significance of the change, and the approximation symbol (≈) indicating no change in the state compared to baseline.

During the stress condition, MeanRR, RMSSD, pNN50, and SD2 maintained their consistency across all UST segments. The trend of SDNN changed at the 30-s segment from higher to lower than baseline. Similarly, total power and SD1 showed an opposite trend at the 120-s segment compared to the 5-min recording. HF power and LF power maintained their trend consistency until the 60-s and 120-s segments, respectively. Conversely, all HRV measures maintained trend consistency during paced breathing. Although the trend direction in pNN50 and SD1 was consistent, the significance of the change varied among the UST segments during paced breathing.

5.6 Discussion

This study investigated the reliability of UST HRV analysis during resting, stress, and paced breathing to provide fundamental considerations for HRV incorporation in real-time systems. The subsequent sections provide a comprehensive discussion of each investigated hypothesis as well as a summary of the overall findings, followed by the limitations of the study.

5.6.1 Minimum Reliable Ultra-Short-Term Segment

Following the recommendations for UST reliability analysis made by Pecchia et al. (2018) and Shaffer et al. (2020), we propose the following validity criteria to examine the reliability of HRV measures derived from UST segments under resting and non-resting conditions, with particular attention to a priori for Bland-Altman bias and the acceptable CI lower bound for Pearson correlation analysis:

- Identification of acceptable mean differences between UST segments and the standard 5-min interval using Bland-Altman analysis with a priori mean difference (i.e., bias) of 10%.
- Establishment of a significant strong relationship between the different time segments and standard 5-min interval, with correlation coefficients greater than 80% ($p < .05$, 95% CI lower bound > 75%).

Further, a supplementary criterion was added for non-resting states to ensure trend consistency compared to baseline:

- Trend consistency in the average value across all time segments from baseline to stress and from baseline to paced breathing conditions, assessed by the direction and significance of the statistical ratio of post-hoc analysis (i.e., t-statistic).

Table 5.4 summarises the results of the minimum reliable UST segment based on the satisfaction of all criteria, while Supplementary Table D.14 outlines the results according to each criterion separately.

Overall, the results of the HRV measures derived from UST analysis during baseline and stress were similar (see Table 5.4). During paced breathing, SDNN met the reliability criteria at the 30-s segment compared to the 60-s segment in baseline and stress. Similarly, LF power during paced breathing was reliable at the 60-s segment, but none of the UST segments were found reliable in the remaining conditions. These findings can be explained by the periodicity of slow HRV oscillations during paced breathing (Russo et al., 2017). Additionally, these ultra-shorter segments (i.e., 30 s, 60 s) included more than one completed respiratory cycle (7 breaths/min), which may have improved the analysis by providing sufficient data regarding the underlying signal pattern (Yu et al., 2018a).

Table 5.4: Proposed Minimum Reliable UST Segment

Measure	Baseline	Stress	Paced Breathing
MeanRR	10	10	10
RMSSD	60	60	60
SDNN	60	60	30
pNN50	120	120	–
LF power	–	–	60
HF power	60	60	120
LF/HF	–	–	–
Total power	–	–	–
SD1	60	–	60
SD2	60	60	60

An important outcome of this study is that a 10-s segment was found to reliably estimate MeanRR in all conditions (i.e., resting, stress, and paced breathing). This finding is consistent with those obtained by Shaffer et al. (2016), who examined UST reliability under resting conditions. Further, we found that a 60-s segment can reliably estimate RMSSD under all conditions. These results are in agreement with those of Esco and Flatt (2014) and Shaffer et al. (2016), which both focused on resting conditions, as well as those of Castaldo et al. (2019), which focused on resting and stress conditions. However, Munoz et al. (2015) conducted a UST reliability study using a large sample size of approximately 3,000 participants and concluded that a 30-s segment, or three averaged 10-s segments, can provide a good estimate of RMSSD. Overall, these results lend further evidence to support recent calls to shorten the window of HRV analysis to less than 5 min to conform to real-time requirements. In particular, the outcomes of this study provide insights related to the minimum reliable segment for HRV analysis in non-resting conditions.

5.6.2 Influence of Condition on Heart Rate Variability

Regarding the influence of stress and paced breathing on the HRV measures, the results of this study are in line with previous findings. For instance, MeanRR significantly decreased during the stress task, indicating an elevation in HR (Khazan, 2013). Although MeanRR approximately returned to baseline during paced breathing, SDNN showed a significant increase, which is consistent with the findings of Lin et al. (2014), Steffen et al. (2021), and You et al. (2021a). Similarly, the frequency-domain measures (e.g., LF power, LF/HF, total power) significantly increased during the paced breathing exercise due to slow respiratory oscillations, which is consistent with the findings of Clamor et al. (2016) and Steffen et al. (2017). The increase in LF power during paced breathing has been shown to be associated with vagal tone by manifesting the RSA (Kromenacker et al., 2018; Shaffer & Meehan, 2020; Van Diest et al., 2014).

Contrary to expectations, the stress task did not produce significant changes in RMSSD and HF power, which are commonly used to index vagal tone at normal breathing rates (Laborde et al., 2017). These results are not consistent with the systematic review conducted by Castaldo et al. (2015), which demonstrated an overall reduction in RMSSD and HF power during mental stress tasks. This phenomenon could be due to the small sample size of the present study or because the designed tasks were not stressful enough, with respect to stress duration or stress type, to significantly decrease parasympathetically related HRV measures.

This study sought to extend investigations into the influence of stress and paced breathing on HRV measures by examining trend consistency in UST segments compared to baseline. During the stress condition, the trend of some of the HRV measures (e.g., SDNN, HF power) changed at segments shorter than 60 s. However, HRV measures maintained trend consistency across all UST segments during paced breathing. These findings provide further considerations for determining the minimum reliable segment for HRV analysis in non-resting conditions.

5.6.3 Overall Discussion

The overall outcomes of this study partially supports H1 in which UST analysis provided reliable estimates for the time-domain HRV measures, HF power and SD2 compared to the 5-min segment under resting, stress, and paced breathing conditions. However, the UST analysis of the LF power was only found to be reliable in the paced breathing condition. Similarly, the results partially support H2 in which HRV measures differed significantly between paced breathing and baseline. However, there was insufficient evidence to conclude that mental stress altered HRV relative to the baseline.

However, these findings may be limited due to the participants' profile because they were recruited via a call for participation made to the computer science department at QMUL. Thus, there is a limit to how broadly the results can be applied to different age groups, academic majors, or occupational stress categories given that demographic factors may influence HRV levels and stress perception.

5.7 Limitations

A significant limitation of this study is the utilisation of a random UST segmentation approach. Although there was an excellent agreement among the various randomised UST segments based on ICC (see Section 5.5.3), extracting a single random UST segment from each RR interval may have presented some drawbacks that could have impaired HRV analysis. A better strategy would be to perform the reliability analysis on multiple extracted UST segments from the 5-minute recording, similar to the strategy adopted by Melo et al. (2018).

Another limitation concerns the method employed to calculate power spectral analysis, which may have resulted in inaccurate frequency-domain measures in the UST segments. This relates to the underlying mathematical assumptions of power spectral analysis regarding the signal length, as discussed by Malik et al., 1996 (see Section 5.2). Additionally, the maximum length of the investigated segment in the present study was 120 s, which may have constrained the reliability analysis of some of the measures (e.g., total power, LF/HF). Lastly, an arguable weakness is the arbitrariness of selecting a predetermined a priori value of 10%. The acceptable a priori threshold should be investigated and defined based on clinical evidence. Future studies could assess the reliability of UST analysis in segments longer than 120 s (e.g., 180 s, 240 s), with particular attention given to the frequency-domain measures by examining other spectral density estimation techniques (e.g., maximum entropy method), as suggested by Shiraishi et al., 2018.

5.8 Chapter Summary

This chapter addresses SRQ2 by providing an empirical investigation of UST HRV analysis for future use in real-time systems and wearable devices under three conditions: resting, stress, and paced breathing. It establishes criterion validity to explore the minimum reliable window for HRV analysis with a standard 5-min interval by assessing correlation, limits of agreement, and trend consistency for non-resting conditions.

The findings of this study indicate that UST reliability assessment differs in paced breathing compared to resting and stress because of differences in signal characteristics under a slow respiratory rate. For instance, SDNN showed high-reliability levels at 30 s during paced breathing compared to 60 s in the other conditions. Based on a priori criteria for Bland-Altman's bias and a Pearson's correlation coefficient of greater than .80 (95% CI lower bound > .75), MeanRR was the only HRV measure that obtained a high agreement level and correlation with the 5-min RR interval at segments less than 1 min across all conditions, with high reliability up to 10 s. At the 60-s segment, RMSSD and SD2 reliably estimated the 5-min recording in all conditions. In contrast, the remaining HRV measures were either reliable at the 120-s segment or not reliable at the examined segments.

Given the notable effects of paced breathing on physiological responses demonstrated by SDNN and LF power, the next chapter further examines the impact of paced breathing on HRV, BP, and a range of affective states, including cognition, relaxation, and stress.

CHAPTER 6

Heart Rate Variability Biofeedback

This chapter examines the short-term effects of heart rate variability biofeedback on mental well-being via a paced breathing exercise. It focuses on the assessment of affective states (attentiveness, fatigue, mood, serenity, and stress), executive function (cognitive performance in a working memory task), and physiological responses (heart rate variability and blood pressure).

6.1 Overview

Around the world, mental health and well-being are widely scrutinised topics; however, greater awareness, improved treatment, and increased support are still needed for both. According to a report from the WHO, the global incidence of mental health conditions (e.g., anxiety, bipolar disorder, depression, neurodevelopmental disorders) was one in every eight people in 2019, with the latest data indicating an increase of more than 20% in 2020 due to the pandemic (WHO, 2022a; 2022b). These findings are crucial as mental health and well-being can have a significant impact on an individual's cognitive and emotional abilities (Dattani et al., 2021). Furthermore, numerous psychiatric disorders can cause abnormal cardiac reactivity, such as increased HR and BP (Alvares et al., 2016).

Therefore, researchers and clinicians have investigated the best methods for regulating HR via increased HRV, which, by extension, increases the parasympathetic activity responsible for stimulating the rest-and-digest response. In general, interaction within the cardiorespiratory system has been shown to improve HRV and RSA, indicating a significant contribution from parasympathetic activity (Vaschillo et al., 2002; Zaccaro et al., 2018). Consequently, HRVB via paced breathing exercises has emerged as a promising non-invasive psychotherapy approach that can effectively build resilience and improve overall health and well-being (Lehrer & Gevirtz, 2014; Steffen et al., 2021).

Many studies have examined the positive impact of HR regulation via long-term HRVB practices on psychophysiological responses, as demonstrated in a meta-analysis of 58 studies conducted by Lehrer et al. (2020). However, relatively little research has investigated the efficacy of single-session, short-term HRVB in terms of instantaneous effects on mental well-being, whether during or after a paced breathing exercise. Hence, this study sought to examine the short-term effects of HRVB on affective states, cognitive performance, and physiological measures, including HRV and BP. The next section reviews the existing literature and presents an associated research question, followed by the study hypotheses.

6.2 Related Work

HRVB plays a significant role in increasing HRV measures, thereby improving physical and mental health (Kemp et al., 2012; Lehrer et al., 2003; Lehrer et al., 2004; Quintana et al., 2012). Lehrer et al. (2013) proposed a simplified protocol for HRVB by determining RF over five sessions. Despite the authors' argument regarding the importance of determining the RF for each individual, more recent research has shown similar physiological behaviour when breathing at a rate of 6 breaths/min (Van Diest et al., 2014; Zaccaro et al., 2018). Preliminary evidence suggests that HRVB stimulates activity in the vagus nerve, which is a major parasympathetic nerve associated with relaxation and reduced stress levels (Gerritsen & Band, 2018).

Several studies have assessed the efficacy of stimulating parasympathetic activity through long-term HRVB interventions (i.e., multiple sessions of paced breathing exercises) to treat mental health conditions, including anxiety (Lagos et al., 2008; Lee et al., 2015), depression (Caldwell & Steffen, 2018; Karavidas et al., 2007), and stress levels (Goessl et al., 2017; Hallman et al., 2011; van der Zwan et al., 2015). However, the short-term impact of a single HRVB session on mental health and cognitive performance remains unclear. In an RCT of a single-session HRVB intervention on musicians' performance, Wells et al. (2012) reported an increase in parasympathetic levels, as reflected by an increase in their HF power. Moreover, the study showed that anxious musicians reported lower anxiety scores after the intervention during a stressful performance.

A similar study was also recently conducted by Laborde et al. (2022), who examined the impact of a single HRVB session on psychophysiological responses. These responses included perceived stress, emotions, and the RMSSD measure from HRV time-domain analysis of two groups performing paced breathing with and without biofeedback. The authors addressed some of the limitations in Wells et al. (2012) by adopting a within-subject research design and recruiting a larger sample size representative of the general population. They reported improvement in emotional control and high levels of RMSSD during the paced breathing exercise. Although Laborde et al.'s (2022) results for parasympathetic activity were similar to those obtained by Wells et al. (2012), the former study lacked a control condition, which hinders interpretation of the results.

Under a theoretical approach, the neurovisceral integration model has been introduced (Thayer & Lane, 2000) and revised (Thayer et al., 2009a) over the last two decades. The proposed model demonstrates the link between the prefrontal cortex and cardiac vagal tone, thus delineating the association between HRV and emotion, cognition, and mental health. For instance, Mather and Thayer (2018) found that HRVB via a paced breathing exercise generated high-amplitude oscillations that influenced brain rhythms, thereby enhancing emotional regulation and well-being.

One of the critical domains of the neurovisceral integration model is the relationship between cognitive function and HRV, particularly RMSSD and HF power (Thayer et al., 2009a; Thayer & Lane, 2000). Forte et al. (2019) identified the major cognitive areas used in HRV studies by conducting a systematic review of the literature and concluded that, in general, higher resting HRV measures were associated with improved cognitive functioning, as examined in tasks assessing attention (Williams et al., 2016), global cognition (Solernó et al., 2012), and memory (Hansen et al., 2004; Hansen et al., 2003).

Accordingly, HRVB emerged as a powerful technique for increasing HRV measures and, by extension, improving cognitive performance. In a seminal empirical study linking HRVB to cognitive function, Prinsloo et al. (2011) demonstrated that a single short-term HRVB session improved cognitive performance, as assessed by the reaction time and response accuracy from a modified Stroop task. However, there is still controversy over the extent of HRVB efficacy on cognitive function, as discussed in a recent systematic literature review by Tinello et al. (2022). While there was a generally positive relationship between HRVB and cognitive performance in the reviewed studies, there was a lack of association between HRV measures and cognition improvement, as half of the reviewed studies (8 out of 16) did not report physiological data following the intervention.

Moreover, research on single-session HRVB interventions has shown that they can increase relaxation (Lin et al., 2020; Prinsloo et al., 2013; Yu et al., 2018a), improve mood (Steffen et al., 2017), and reduce perceived stress and anxiety (Kennedy & Parker, 2019; Lemaire et al., 2011; Schuman & Killian, 2019; Yu et al., 2018b). In particular, Steffen et al. (2017) conducted a controlled study to examine the effect of a single HRVB session on BP, HRV physiological measures, and mood in three groups: 1) RF-breathing at the individual's determined RF, 2) RF+1-breathing at 1 breath/min higher than the determined RF, and 3) breathing at a normal rate acting as a control. They found that the RF group reported a higher positive mood and lower SBP in response to the stress task than the other two groups. Nonetheless, both paced breathing groups (RF and RF+1) showed improvements in the physiological measures. Overall, this study presents promising findings regarding the benefits of a short-term HRVB intervention on emotional and physiological responses.

A few studies have addressed the short-term lasting effects of a single HRVB session on physiological responses. Using a within-subject design to study 24 athletes, You et al. (2021a) examined the influence of a paced breathing exercise without a biofeedback element on the RMSSD measure during, immediately after, and 1 h after the intervention. They found a significant increase in RMSSD during the intervention, which returned to baseline post-intervention. Similarly, Laborde et al. (2019b) investigated the impact of paced breathing on executive function in athletes after physical exercise; however, they incorporated a larger sample size of 107 in a mixed-factorial design. The authors utilised the Stroop task as a stress inducer and to assess inhibitory control (see Section 2.1.2). Although the paced breathing group showed significant improvements in perceived stress and cognitive performance, the improvement in cognitive performance was not mediated by vagal tone, as measured by RMSSD.

Due to the notable paucity of evidence-based literature addressing the potential impact of a single-session, short-term HRVB intervention, the present study aimed to investigate the influence of HRVB on a range of affective states (attentiveness, fatigue, mood, serenity, and stress), executive function (cognitive performance in a working memory task), and physiological measures (HRV and BP). Moreover, the HRV measures during and after the paced breathing exercise were investigated as predictors of the individual participants' affective states via the following research question:

> **SRQ3:** What is the effect of a single paced breathing session on affective states (cognition, relaxation, stress) and physiological responses (HRV and BP)?

6.3 Hypotheses

The hypotheses associated with SRQ3 are as follows:

> **Hypothesis 1 (H1):** HRVB improves **affective states**: attentiveness, fatigue, mood, serenity, and stress.

> **Hypothesis 2 (H2):** HRVB improves **cognitive performance** as demonstrated by working memory.

> **Hypothesis 3 (H3):** HRVB results in lower **BP** reactivity during a cognitive stress task.

> **Hypothesis 4 (H4):** Affective states and cognitive performance can be predicted from HRV measures during and after a paced breathing exercise.

6.4 Methods

6.4.1 Participants

The sample size was determined based on a priori power analysis; thus, the target sample size was estimated as a total of 34 participants using G*Power for a repeated measures ANOVA with a between-subjects factor design, statistical power (1-β) of 80%, significance level (α) of .05, correlation among repeated measures of .50, and capacity to detect a large effect size (f) of .40 (Erdfelder et al., 1996).

Accordingly, 44 healthy adults (aged 23-62 years) were recruited to participate in this study. They were randomly assigned to either an intervention or control group. Participants were recruited from HBKU in Qatar as well as the general community through email advertisements and personal invitations. The exclusion criteria included physical health conditions related to cardiovascular or respiratory diseases, diagnosed psychiatric conditions, and an age outside the range of 18-65 years at the time of recruitment. Participants were asked to avoid caffeine, smoking, and eating heavy meals for 2 h prior to the study as well as engaging in intense physical workouts for 24 h, to minimise any confounding effects on the physiological responses (Laborde et al., 2017; Quintana et al., 2016b). The study was approved by the IRB at Qatar Biomedical Research Institute at HBKU (QBRI-IRB-2021-03-088), as data collection was conducted in the State of Qatar (see Section 3.3.1). All participants were informed about the nature of the experiment, and they signed a digital consent form. All documents related to ethical approval and consent are shown in Appendix E (see Figures E.1 and E.2).

After data collection, the HRV recordings were visually inspected and filtered for noise and artefacts. Consequently, the data of six participants were discarded due to poor signal quality, where noise exceeded 5% of the recording. Out of the remaining 38 participants, 20 were women (mean age: 35.5 $\pm$ 11) and 18 were men (mean age: 34.4 $\pm$ 9.86).

6.4.2 Experiment Design

An RCT study was designed to investigate the impact of HRVB through paced breathing on a range of affective states, executive function, and physiological measures in healthy individuals. The study was based on a mixed-factorial design with two independent variables: group (i.e., between-subjects) and time (i.e., within-subject). Participants were randomly assigned to one of two groups: 1) paced breathing with a biofeedback intervention (HRVB) or 2) normal breathing with no intervention, acting as a control group (CTRL).

The HRV data and affective state questionnaires were collected at four-time points during the study: 1) baseline, 2) pre-intervention (i.e., during the first cognitive task), 3) mid-intervention, and 4) post-intervention (i.e., during the second cognitive task). To measure BP response to the cognitive stress task, the BP data were only collected at baseline, pre-intervention, and post-intervention (see Figure 6.1). Similar to the previous study discussed in Chapter 5, the HRV data were collected using a PPG-based sensor attached to the non-dominant hand. The SBP and DBP measurements were acquired through an upper arm cuff device.

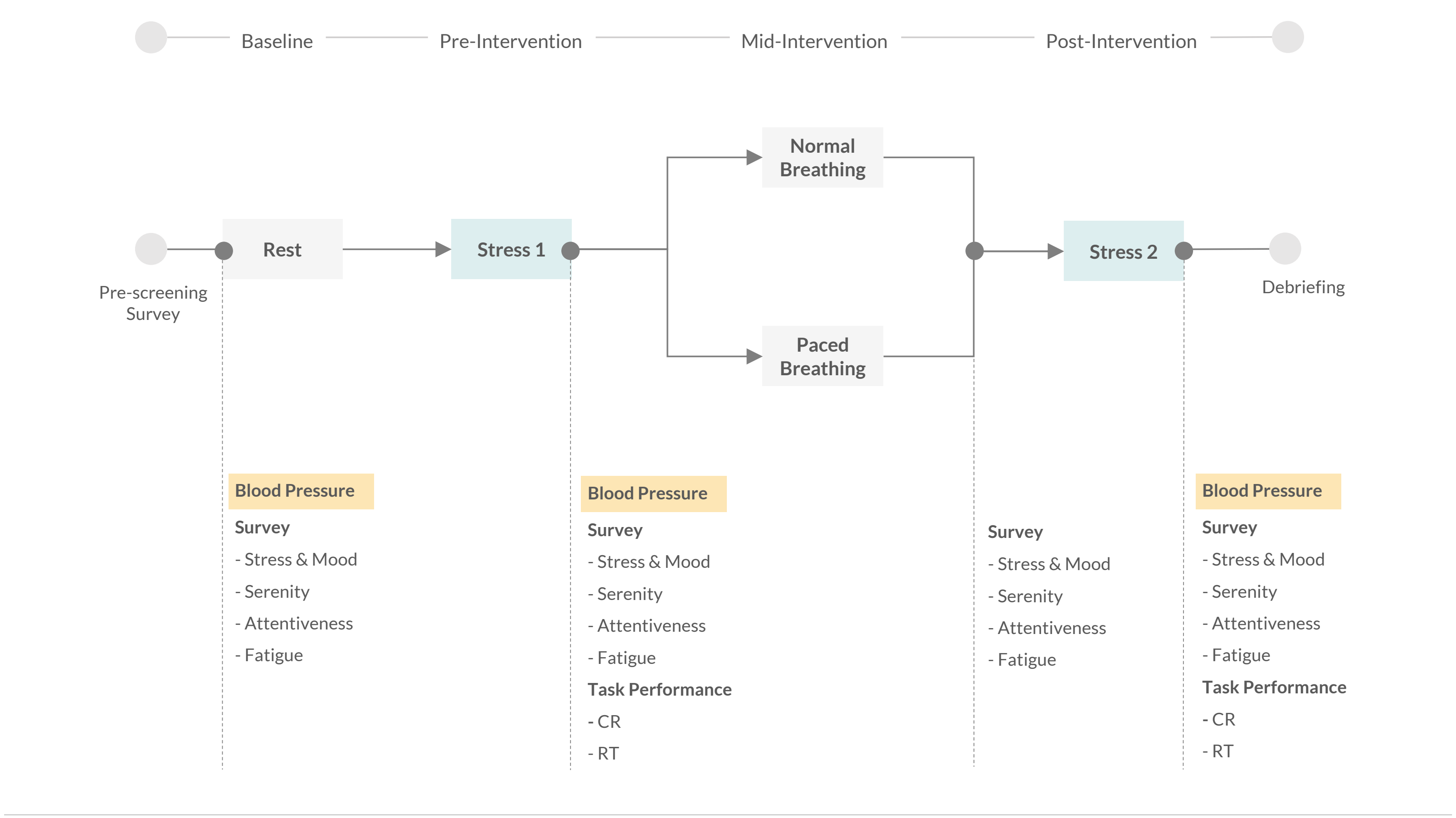


Figure 6.1: A Flowchart for the Experimental Protocol.
Note. HRV was collected during baseline, pre-, mid-, and post-intervention.

Questionnaires

Generally, the study requirements were similar to those of the previous study discussed in Chapter 5. Hence, all participants also filled out the HRV-related questionnaire developed by Quintana et al. (2016b) as a screening survey to assess their eligibility. Furthermore, participants completed a set of questionnaires during baseline, including questions related to demographics; PANAS to assess emotional state; DASS-21 to assess depression, anxiety, and stress; IPAQ to measure physical activity; and PSQI to assess sleep quality (see Section 3.3.2 and Appendix A).

In addition, affective states were measured during the experiment via self-reported questionnaires on mood and stress using a single-item Likert scale, while attentiveness, fatigue, and serenity were measured using the PANAS-X, which is based on a multiple-item Likert scale (see Appendix A.8).

N-Back Task

Although executive function can be considered a component of affective state, it was assessed separately in this study using the N-back task, a computer-based cognitive test (Kirchner, 1958).

The N-back test was employed to assess the cognitive performance of the working memory capacity and serve as a stress inducer. It consists of a series of random alphabetical letters presented on a laptop screen. Participants had to determine whether the current letter was the same as the letter presented in the prior N trials. For this study, the N-back task was implemented using the PsyToolkit web-based platform, and N was set to 2 (Stoet, 2010; 2017; see Appendix E). Each letter was presented for 1800 ms, followed by a 500-ms blank screen period (see Figure 6.2). If there was a match, the participant had to press "M" on the keyboard; otherwise, no response was required. Hence, there were three possible responses to this cognitive task: correct responses (CR), missed responses (MR), and false alarms (FA), as explained in Table 6.1.

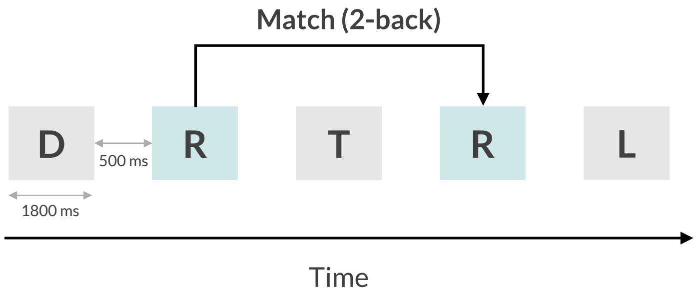


Figure 6.2: A Schematic Representation of the 2-Back Task

reaction time (RT) was measured as the amount of time it took for the participant to correctly press the letter "M" on the keyboard after the letter was presented. The task consisted of two blocks, each with 50 trials. All letters were randomised to eliminate the possibility of bias. Participants also underwent a training session consisting of 25 trials to become familiar with the task prior to completing it. To measure cognitive performance improvement, participants had to solve the N-back task at two-time points: pre- and post-intervention.

Table 6.1: Possible Responses to the N-Back Task

Correct Responses	Participant correctly pressed "M", where the current letter matched the letter from two trials ago.
Missed Responses	Participant did not press "M", although the current letter matched the letter from two trials ago.
False Alarms	Participant incorrectly pressed "M", where the current letter did not match the letter from two trials ago.

Physiological Data

After each cognitive task, BP was measured for all participants by wrapping the cuff of an OMRON M7 Intelli IT around the upper arm. HRV was recorded using CorSense by Elite HRV as this device could be easily attached to the finger, with a sampling rate of 500 Hz. Each measurement was preceded by a 20-s stabilisation period to allow HR to level out. In the event of any technical issues, HRV was measured for 6 min to ensure a minimum recording length of 5 min. Participants were instructed to minimise hand movements as much as possible to maintain a high-quality signal. The Systole Python package (Legrand & Allen, 2022) was used for the HRV time- and frequency-domain analysis, while the pyHRV Python package (Gomes, 2018) was used for the non-linear methods. All signals were filtered using the adaptive threshold artefact detection and moving window average correction methods, based on the approach discussed in Chapter 4.

Data Analysis Approach

For this experimental study, the statistical mean differences between the groups were examined for H1-H3 using ANCOVA, with particular attention given to the effect of other covariates (e.g., baseline measurements, pre-intervention measurements). An ANCOVA is essentially an ANOVA of adjusted dependent variables after controlling for the effect of covariates (Bonate, 2000). The independent variables were set with the group factor (HRVB vs. CTRL; between-subject) and time factor (baseline, pre-, mid-, post-intervention; within-subject). In contrast, the dependent variables consisted of psychophysiological and cognitive measures collected during the experiment, including attentiveness, fatigue, and serenity scores; CR and RT retrieved from the cognitive task; and BP and log-transformed HRV measures. Additionally, MLL analysis was conducted on HRV measures to assess group mean differences, followed by Tukey's test for post-hoc analysis of pairwise comparisons. Lastly, logistic and linear regression analyses were used for H4 to examine the relationship between HRV and cognitive performance as well as the relationship between HRV and affective states.

As a measure of effect size, the omega squared (ω^2) was reported for all ANOVA tests and their variations, including ANCOVAs and MLL analysis. The values of .01, .06, and .14 were interpreted as small, medium, and large effect sizes, respectively. Moreover, Hedges' g was reported for all independent pairwise comparisons, with the values of .2, .5, and .8, indicating small, medium, and large effect sizes, respectively. For the dependent pairwise comparisons, the correlation coefficient (r) was reported with the values of .1, .3, and .5, indicating small, medium, and large effect sizes, respectively (Cohen, 1988).

6.4.3 Procedure

The experimental sessions were conducted during the daytime (9:00 am to 2:00 pm) in the HCI lab at HBKU, and each session lasted 45 minutes. The HCI lab is a quiet small room designed to facilitate in-person experimental studies (see Figure E.3). The experiment protocol involved collecting psychophysiological data at 1) baseline, 2) pre-intervention, 3) mid-intervention, and 4) post-intervention.

Upon arrival at the lab, participants were asked to sign the consent form and fill out the baseline questionnaires related to demographic information as well as anxiety, depression, stress, emotional state, physical fitness, and sleep quality. Subsequently, a 6-min HRV recording was collected at rest, followed by a BP measurement. Following the baseline measurements, participants began the first cognitive task, which was presented on a laptop screen; HRV data were also recorded during this period. The participants were left alone in the room to complete the cognitive tasks. However, the researcher was present between phases to ensure there were no technical issues and to address any concerns. Subsequently, participants received a randomly generated message on the screen indicating their group assignment to Group 1 or Group 2 (CTRL or HRVB, respectively). The HRVB group had to perform a paced breathing exercise by following a breathing guide for 6 min using the Elite HRV deep breathing feature, which shows an opening and closing circle on an iPad screen. Additionally, a biofeedback element was conveyed to participants via visualisation of the sinusoidal wave of the HRV signal during breathing, along with instructional audio prompting inhalation and exhalation.

A brief explanation of the trend between HRV and breathing was communicated to participants so that they would be aware of any deviations from the expected sinusoidal pattern. For the paced breathing activity, a 2-min training session was conducted to ensure that participants were able to perform the breathing activity correctly. The breathing ratio was set to 4 s of inhalation and 6 s of exhalation (6 breaths/min). Participants in the CTRL were asked to sit quietly for 6 min and breathe normally (i.e., similar to baseline). Lastly, all participants had to solve the second cognitive task. Figure 6.3 shows the experiment setup, including the CorSense sensor, BP device, the biofeedback interface, and the cognitive task screen.

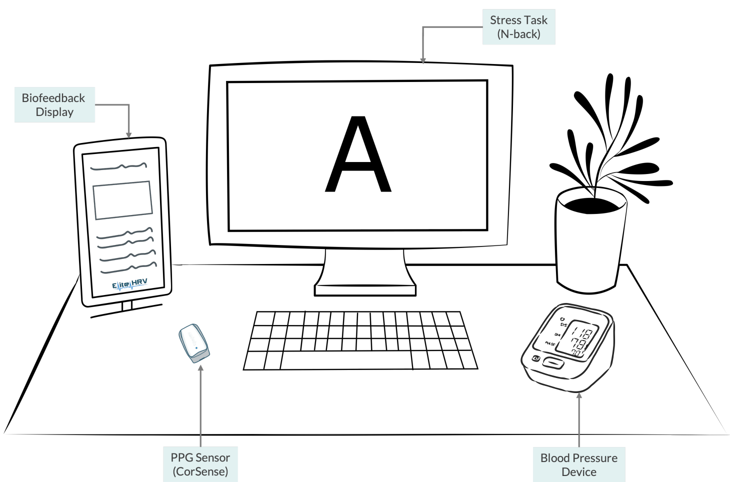


Figure 6.3: An Illustration of the Experimental Setup

6.5 Data Analysis

The preliminary analysis assessed by histogram plots and the Shapiro-Wilk test showed that the HRV measures were not normally distributed ($p < .05$); thus, the data were logarithmically transformed. Hence, all parametric statistical analyses were performed on the log-transformed HRV measures to obtain a better approximation of the normal distribution. In this chapter, the HRV data are presented in their non-transformed form (i.e., absolute values) for simplicity and ease of interpretation, while the log-transformed data are included in Appendix E (see Tables E.1-E.4).

6.5.1 Descriptive Statistics

A descriptive statistical overview of demographic characteristics and baseline psychophysiological measures based on the group are shown in Table 6.2. At baseline, there were no significant differences between the groups in terms of age; BP; BMI; depression, anxiety, and stress (DASS-21); experience with deep breathing; experience with meditation; HRV measures; physical activity level (PSQI); positive and negative affective states (PANAS-X); or sleep quality (PSQI; see Table 6.2). Gender was balanced, with 10 women and nine men in each group. The effect of gender on the baseline HRV measures was also examined, and the independent t-test revealed no significant mean differences (p-values > .32 for all; see Figure 6.4).

The correlations among variables presented in Table 6.2 were calculated using the Pearson's correlation coefficient of the log-transformed HRV measures. There were significant positive relationships between the negative affective states measured by the PANAS and DASS-21 (depression: r = .69, anxiety: r = .72, stress: r = .71, $p < .05$) and sleep quality measured by the PSQI (r = .39, $p < .05$). As expected, there were significant inverse relationships between age and the

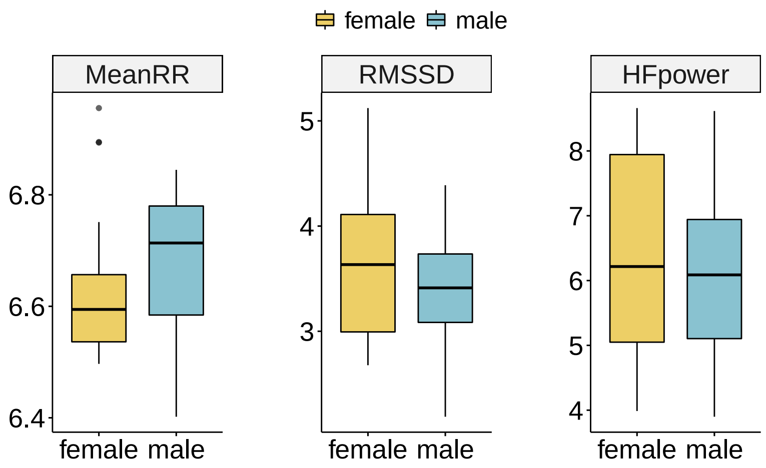


Figure 6.4: Boxplots for the Baseline MeanRR, RMSSD, and HF Power based on Gender.

baseline HRV measures, including SDNN, SD1, and SD2 (Laborde et al., 2017; Quintana et al., 2016b). However, there were no significant relationships between age and the remaining HRV measures. Figure 6.5 demonstrates the relationship between age and baseline SDNN based on gender in a scatter-density plot.

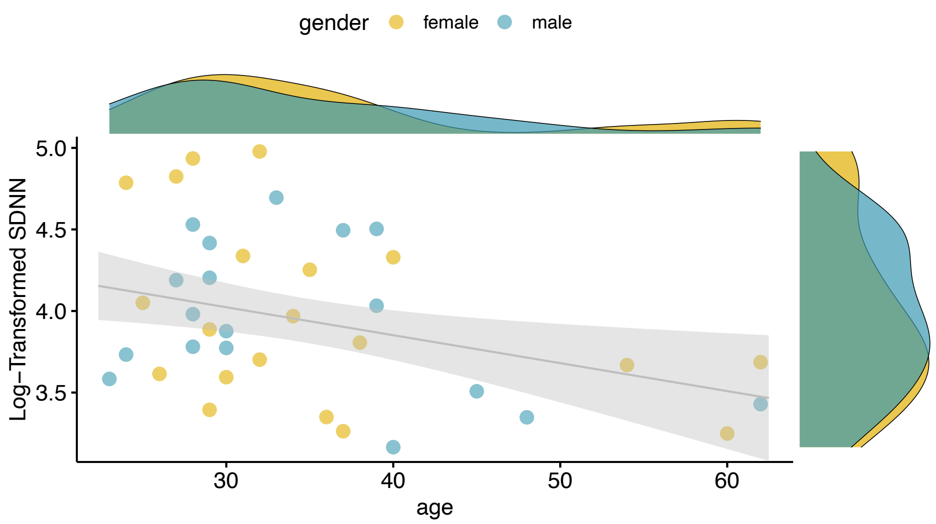


Figure 6.5: Scatterplot and Density plots for SDNN based on Age and Gender

Table 6.2: Demographic and Baseline Characteristics by Group (N=38)

	Total mean (SD)	CTRL mean (SD)	HRVB mean (SD)	t	p
Demographic					
Age	34.9 (10.3)	33.6 (6)	36.3 (13.4)	-0.78	.44
BMI	25.7 (4.7)	25.6 (5.1)	25.8 (4.3)	-0.13	.90
Deep Breathing	2.5 (0.95)	2.63 (1.0)	2.37 (0.90)	0.85	.40
Meditation	2.08 (1.0)	2.21 (1.1)	1.95 (0.91)	0.81	.42
PANAS					
Positive mood	24.4 (7.9)	26.2 (6.6)	22.5 (8.9)	1.45	.16
Negative mood	8.95 (6.5)	8.7 (5.9)	9.2 (7.2)	-0.20	.84
DASS-21					
Depression	1.58 (1.5)	1.63 (1.5)	1.53 (1.5)	0.22	.83
Anxiety	2.2 (1.6)	2.2 (1.5)	2.1 (1.5)	0.21	.84
Stress	1.1 (1.4)	1.1 (1.4)	1.1 (1.4)	0.00	1.0
PSQI					
Sleep Quality	6.66 (3.4)	7.21 (3.7)	6.11 (3.1)	1.0	.32
IPAQ					
MET	1547 (2196)	1092.3 (1152)	2002 (2855.3)	-1.29	.21
kCal	1848 (2765)	1198.8 (1068.3)	2497.3 (3699.5)	-1.47	.16
HRV Time					
MeanRR	783 (104)	766 (84)	800 (121)	-1.0	.32
RMSSD	46 (49)	52 (44)	40 (33)	0.97	.34
SDNN	60.4 (33)	61 (36)	59 (31)	0.18	.86
pNN50	20.4 (20)	23 (21)	18 (20)	0.80	.43
HRV Frequency					
LF Power	1557 (2799)	1949 (3277)	1165 (2246)	0.86	.40
HF Power	1266 (1707)	1368 (1856)	1164 (1588)	0.36	.72
LF/HF	1.8 (1.6)	2.1 (1.9)	1.6 (1.3)	0.97	.34
Total power	3691 (4992)	4308 (5820)	3075 (4068)	0.76	.45
HRV Non-Linear					
SD1	42.9 (34.6)	44 (36)	42 (34)	0.13	.90
SD2	72.4 (34.7)	74 (39)	71 (31)	0.25	.81
Blood Pressure					
Systolic	102.8 (13.8)	103.1 (13.7)	102.4 (14.2)	0.15	.88
Diastolic	76.3 (6.0)	77.5 (5.6)	75.0 (6.3)	1.32	.20

6.5.2 H1 | Affective States

The first hypothesis posed that a single session of HRVB would have positive effects on levels of perceived attentiveness, fatigue, mood, serenity, and stress.

Methods

Affective states were measured subjectively based on self-reported questionnaires of attentiveness, fatigue, mood, serenity, and stress at four-time points: 1) baseline, 2) pre-, 3) mid-, and 4) post-intervention.

Participants had to rate their stress and mood on a five-point single-item Likert scale: for stress, a score of 1 indicated "not at all stressed", while 5 indicated "extremely stressed"; for mood, 1 indicated "extremely sad", while 5 indicated "extremely happy". Attentiveness, fatigue, and serenity components were adopted from the PANAS-X self-report questionnaire (see Appendix A.8). Each component consisted of multiple terms measuring affective states at that moment, and these terms had to be rated 1 for "not at all" to 5 for "extremely". Table 6.3 lists the terms for each assessed component.

Table 6.3: PANAS-X Terms

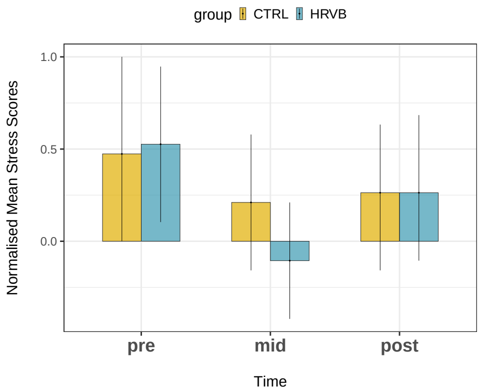
Attentiveness	alert, attentive, concentrating, determined
Fatigue	drowsy, sleepy, sluggish, tired
Serenity	at ease, calm, relaxed

As the data were collected from single-item Likert scale questions, stress and mood were ordinal metrics. The remaining PANAS-X components were interval-level metrics because scores were calculated as the sum of multiple Likert scale questions. Therefore, mood and stress were analysed based on standard descriptive statistics by calculating the change in scores from baseline, and the PANAS-X components were analysed using inferential statistics of the group mean differences. The outcomes were labelled as low or high based on the differences between the reported scores from the mid- and post-intervention time points (independently) and the reported scores from the pre-intervention time point. The CTRL group and low scores were coded as 0, and the HRVB group and high scores were coded as 1. The relationship between both groups (CTRL vs. HRVB) and the PANAS-X components was assessed using Spearman's rank-order correlation (r_s).

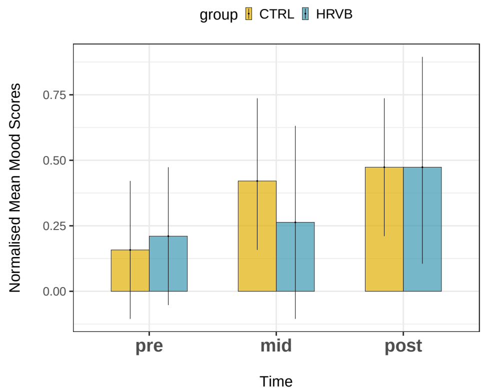
Results | Stress and Mood

After the first cognitive task (i.e., pre-intervention), participants in both groups reported higher average stress scores compared to baseline. However, the mid-intervention average stress score of the HRVB group (M = -.12, SD = .73) was lower than the average stress score of the CTRL (M = .21, SD = .86), as shown in Figure 6.6a. This indicates that the participants in the HRVB group reported lower perceived stress levels immediately after the intervention in comparison to the CTRL. However, both groups reported a similar stress score average (M_{both} = .26, SD_{HRVB} = .93, SD_{CTRL} = .87) after the second cognitive task (post-intervention).

Regarding the mood scores, participants in both groups reported approximately similar average scores pre-intervention (M_{CTRL} = .16, SD_{CTRL} = .60; M_{HRVB} = .21, SD_{CTRL} = .54). Although the average mood scores slightly improved as the experiment progressed, the CTRL (M = .42, SD = .61) reported higher average mood scores mid-intervention than the HRVB (M = .26, SD = .81), as shown in Figure 6.6b. However, there was overlap in all CIs for the mood and stress scores (see Figure 6.6), indicating a lack of significant between-group differences at each time point.



(a) Stress



(b) Mood

Figure 6.6: Average Changes in Stress and Mood Scores at each Time Point from Baseline

Results | Attentiveness, Fatigue, and Serenity

While controlling for the effect of the pre-intervention scores, a two-way mixed ANCOVA was performed to assess the impact of the group (HRVB vs. CTRL; between-subject) and time (mid- and post-intervention; within-subject) on the affective states reported from the PANAS-X questionnaire, including attentiveness, fatigue, and serenity scores. All scores at the pre-, mid-, and post-intervention time points were normalised by subtracting the scores reported at baseline (see Figures 6.7-6.9). Skewed data were log-transformed, and the Shapiro-Wilk test was conducted to ensure normality ($p > .05$ for all). Homogeneity of variances was found for all PANAS-X components, as assessed by Levene's test ($p > .36$ for all). A visual inspection of the scatterplots revealed a linear relationship between the covariates and dependent variables for each group. Moreover, there was homogeneity of the regression slopes as the interaction term was not statistically significant for attentiveness ($F(1, 68) = 2.16$, $p = .15$) fatigue ($F(1, 68) = .01$, $p = .93$), and serenity ($F(1, 68) = .02$, $p = .88$). Pre-intervention covariate scores were significantly related to mid- and post- intervention scores: (serenity: $F(1, 71) = 2.25$, $p < .05$, $r = .26$, attentiveness:

$F(1,71) = 12.56$, $p < .05$, $r = .75$, fatigue: $F(1,71) = 9.51$, $p < .05$, $r = .83$).

The attentiveness score had a significant effect on the group factor after controlling for the effect of covariates ($F(1,71) = 12.57$, $p < .0001$, $\omega^2 = .17$). Planned contrasts revealed a significant difference between the HRVB and CTRL groups at the mid- ($t(71) = 2.74$, $p < .05$, Hedges' $g = .79$) and post-intervention time points ($t(71) = 3.05$, $p < .05$, Hedges' $g = .89$). The time factor did not have a significant effect; however, participants in the HRVB and CTRL groups reported a higher attentiveness score after the second cognitive task (i.e., post-intervention) in comparison to the previous time point ($M_{HRVB} = .73$, $M_{CTRL} = .53$, SE = .47). Furthermore, the group factor (HRVB or CTRL) was significantly correlated to changes in the attentiveness scores from pre- to mid-intervention ($r_s = .36$, $p = .03$) and pre- to post-intervention ($r_s = .41$, $p = .01$).

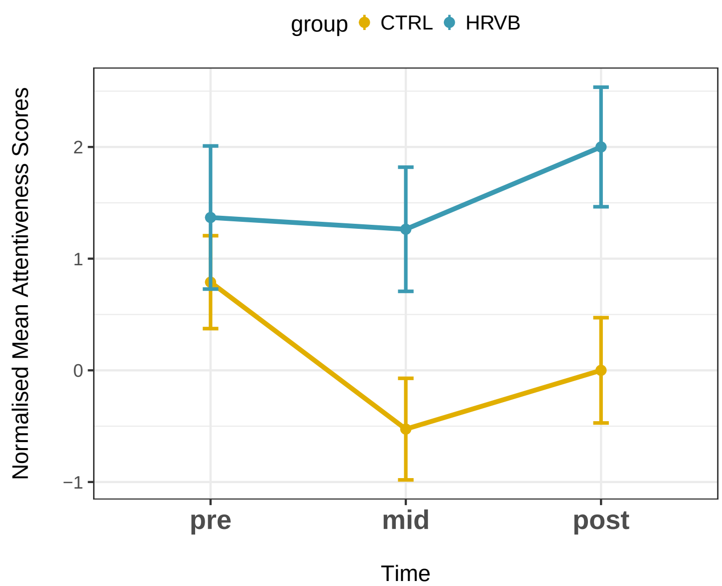


Figure 6.7: Average Changes in Attentiveness Scores at each Time Point from Baseline

Regarding the fatigue scores, there were no significant differences between the groups or across the mid- and post-intervention time points ($p > .05$). However, participants in the HRVB group reported a higher average difference at mid-intervention (M_{adj} = .95, SE = .98) than the CTRL group (M_{adj} = .47, SE = .69), representing a small-sized effect (Hedges' g = .13). Conversely, the HRVB group reported a lower average difference in fatigue scores after the second cognitive task in comparison to the CTRL group (M_{adj} = -.54, SE = .73), representing a small-sized effect (Hedges' g = .19). In addition, there was a non-significant correlation between the group factor and changes in fatigue scores from the pre- to mid-intervention ($r_s = .16$, $p = .30$) and pre- to post-intervention ($r_s = .17$, $p = .30$) time points.

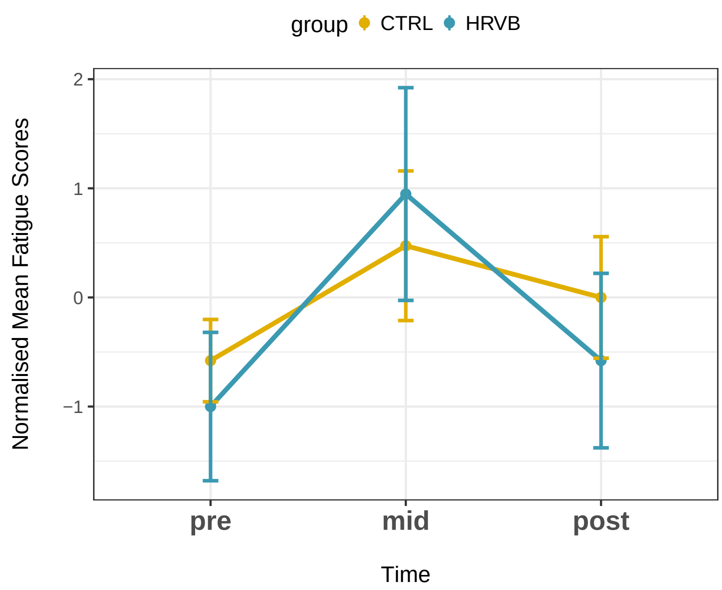


Figure 6.8: Average Changes in Fatigue Scores at each Time Point from Baseline

The serenity scores had a significant effect on the group factor after controlling for the effect of the covariates, ($F(1,71) = 25.18$, $p < .0001$, $\omega^2 = .37$). Planned contrasts revealed that the paced breathing exercise significantly increased the serenity scores mid- ($t(71) = 5.02$, $p < .0001$, Hedges' $g = 1.66$) and post-intervention ($t(71) = 4.47$, $p < .0001$, Hedges' $g = 1.35$). Although there were no significant effects of the time factor, the average adjusted serenity score decreased after the second cognitive task (post-intervention) in comparison to the mid-intervention time point for the HRVB group ($M_{diff} = -.97, t = -1.24, p = .60, r = .15$) and the CTRL group ($M_{diff} = -.53, t = .69, p = .90, r = .08$). Moreover, the group factor was significantly correlated with the changes in serenity scores from pre- to mid-intervention ($r_s = .63, p < .001$) and pre- to post-intervention ($r_s = .44, p = .006$).

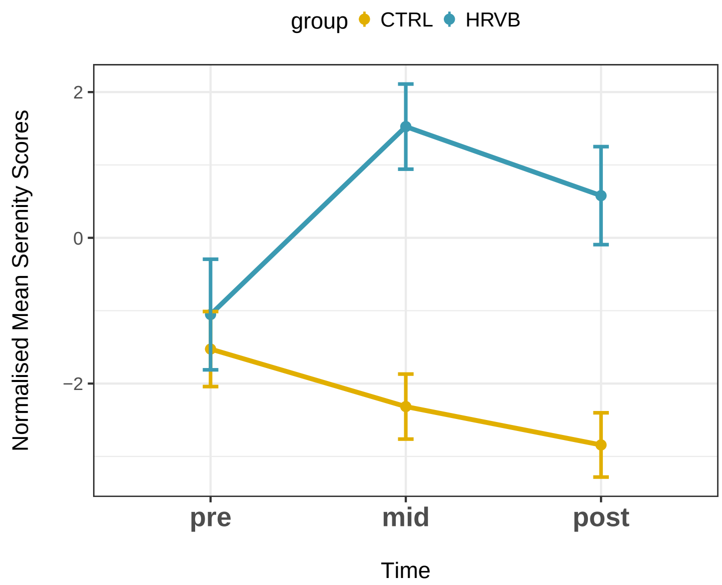


Figure 6.9: Average Changes in Serenity Scores at each Time Point from Baseline

6.5.3 H2 | Executive Function

The second hypothesis stated that a single session of HRVB would improve the cognitive performance via a working memory task.

Methods

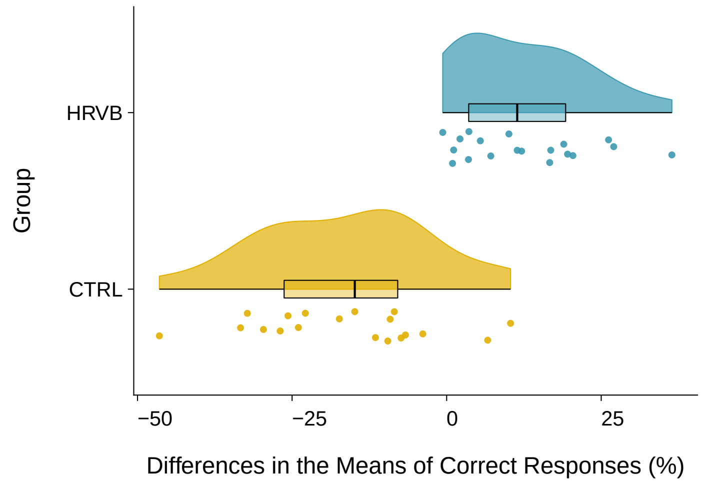
Cognitive performance was assessed using the N-back task by evaluating three metrics: CR, FA, and RT. Participants were asked to solve the task pre- and post-intervention. Hence, a one-way ANCOVA was run to assess group differences in the post-intervention task by controlling for the effect of performance on the pre-intervention task. The Shapiro-Wilk test was conducted to ensure normality on all cognitive performance metrics for each group. The results were non-significant, indicating that the metrics were approximately normally distributed ($p > .11$ for all). The variance was homogeneous between the groups for all metrics, as assessed by Levene's test ($p > .56$ for all).

In addition, the correlations between group (CTRL, coded as 0; HRVB, coded as 1) and cognitive performance were assessed using Spearman's rank-order correlation (r_s). Cognitive performance was represented as a dichotomous variable with low and high outcomes (coded as 0 and 1, respectively) based on the difference in CR between pre- and post-intervention scores.

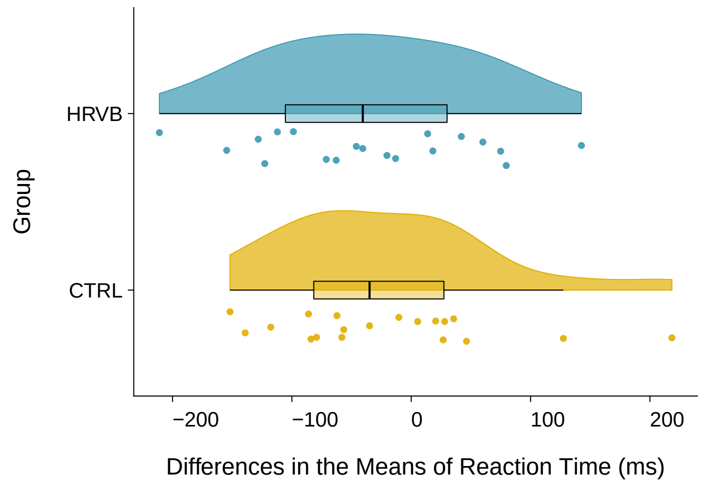
Results

The preliminary analysis for both groups showed a linear relationship between pre- and post-intervention results for CR, FA, and RT, as assessed by visual inspection of the scatter plots. There was homogeneity of the regression slopes given that the interaction was not statistically significant for CR ($F(1, 34) = 2.56,\ p = .11$), FA ($F(1, 34) = .16,\ p = .69$) and RT ($F(1, 34) = .28,\ p = .60$). Further, the covariates were significantly related to the post-intervention results for CR ($F(1, 35) = 30.06,\ p < .0001,\ r = .98$), FA: ($F(1, 35) = 34.56,\ p < .0001,\ r = .99$), and RT ($F(1, 35) = 9.72,\ p < .005,\ r = .98$).

After adjusting for the pre-intervention scores, a significant difference in the CR results was found between the HRVB and CTRL groups ($F(1, 35) = 81.55,\ p < .0001,\ \omega^2 = .68$). Tukey's post-hoc analysis showed a significant increase in CR in the HRVB group (M_{diff} = 29, SE = 2.3, $t(35) = 9,\ p < .0001$). In contrast, the average FA post-intervention was not significant between the HRVB and CTRL groups (M_{diff} = .4, SE = .93, $F(1, 35) = .08,\ p = .78$). Further, the Spearman's rank-order correlation showed a statistically significant relationship between group and CR ($r_s = .84,\ p < .001$). Regarding the RT, there was no significant difference between both groups ($F(1, 35) = .10,\ p = .77,\ \omega^2 = -.02,\ r_s = -.05$). Figure 6.10 shows density and box plots for the differences in CR and RT.



(a) Correct Responses



(b) Reaction Time

Figure 6.10: Results of the Difference in Cognitive Performance (Post-Pre)

6.5.4 H3 | Blood Pressure

The third hypothesis posited that a single HRVB session would result in lower BP reactivity during the cognitive stress task.

Methods

BP was measured as SBP and DBP, with the measurements taken at baseline, pre-intervention, and post-intervention time points. A one-way ANCOVA was applied to analyse the differences in BP between the CTRL and HRVB groups post-intervention, with a covariate of BP measurements collected pre-intervention. Subsequently, Tukey's post-hoc analysis was performed, and the Shapiro-Wilk test was conducted to ensure normality. All SBP and DBP measurements were non-significant, indicating that the measurements were approximately normally distributed ($p > .14$ for all). The variance was homogeneous between both groups, as assessed by Levene's test ($p > .17$ for all).

Results

The preliminary analysis showed a linear relationship between the pre- and post-intervention SBP and DBP measurements for both groups, as assessed by a visual inspection of the scatterplots. The interaction term was not statistically significant for SBP and DBP ($p > .06$); thus, there was homogeneity in the regression slopes. The pre-intervention SBP and DBP measurements were significantly related to post-intervention BP measurements (SBP: $F(1, 35) = 104.5,\ p < .0001,\ r = 1$; DBP:$F(1, 35) = 83.17,\ p < .0001,\ r = 1$).

After adjusting for the pre-intervention BP measurements, a statistically significant difference in post-intervention SBP was found between the HRVB and CTRL groups ($F(1, 35) = 46.2,\ p < .0001,\ \omega^2 = .54$). Tukey's post-hoc analysis revealed a significant decrease in SBP in the HRVB group compared to the CTRL group (M_{diff} = -16 mmHg, $t(35) = -6.80$, $p < .0001$, $r = .75$). However, there was no significant difference in the DBP measurements between the groups ($F(1, 35) = .2$, $p = .66$ $\omega^2 = .02$). Figure 6.11 shows the average SBP and DPB measurements at baseline, pre-, and post-intervention.

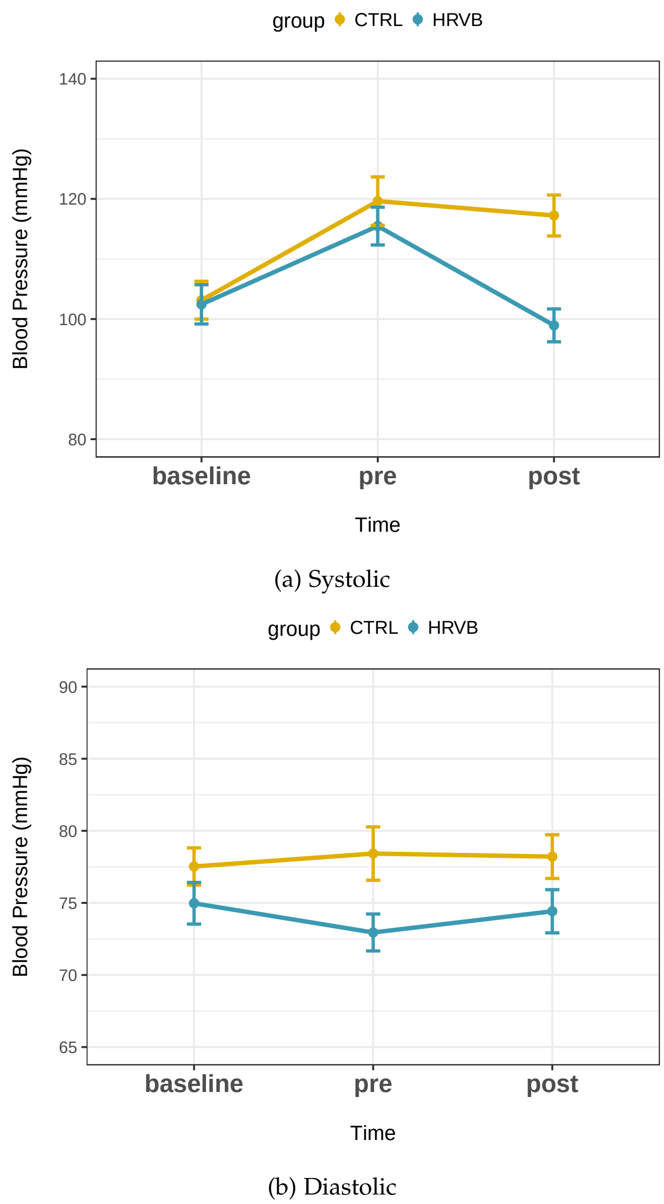


Figure 6.11: Blood Pressure Measurements

6.5.5 H4 | Heart Rate Variability Measures

The primary aim of the fourth hypothesis was to investigate whether the examined affective states and cognitive performance could be predicted from HRV. Based on the results obtained from Section 6.5.2, serenity level was selected as the affective state for examination in this analysis. Accordingly, two major statistical analyses were conducted. First, the relationship and group mean differences were determined between underlying physiological responses, as indexed by HRV, and affective states, as measured by serenity scores and cognitive performance. Second, regression analyses were conducted to determine if the HRV measures could be used to predict serenity and cognition levels.

Impact of Condition on Heart Rate Variability Measures

Taking age into consideration as a covariate, an MLL mixed model was employed to analyse the HRV measures based on the within-subject time factor (baseline, pre-, mid-, and post-intervention) as well as between the groups (CTRL vs. HRVB). This model was used instead of a two-way mixed ANOVA due to a violation of the sphericity assumption, as assessed with Mauchly's test ($W = .49,\ p < .05, \epsilon = .67$). To perform pairwise comparisons, Tukey's post-hoc analysis was subsequently conducted for each HRV measure. To ensure normality, the Shapiro-Wilk test was calculated for the log-transformed values of each experimental condition. The results across all conditions were non-significant, indicating that the measurements were normally distributed ($p > .05$ for all). Figure 6.12 illustrates the average HRV measures at each condition for CTRL and HRVB.

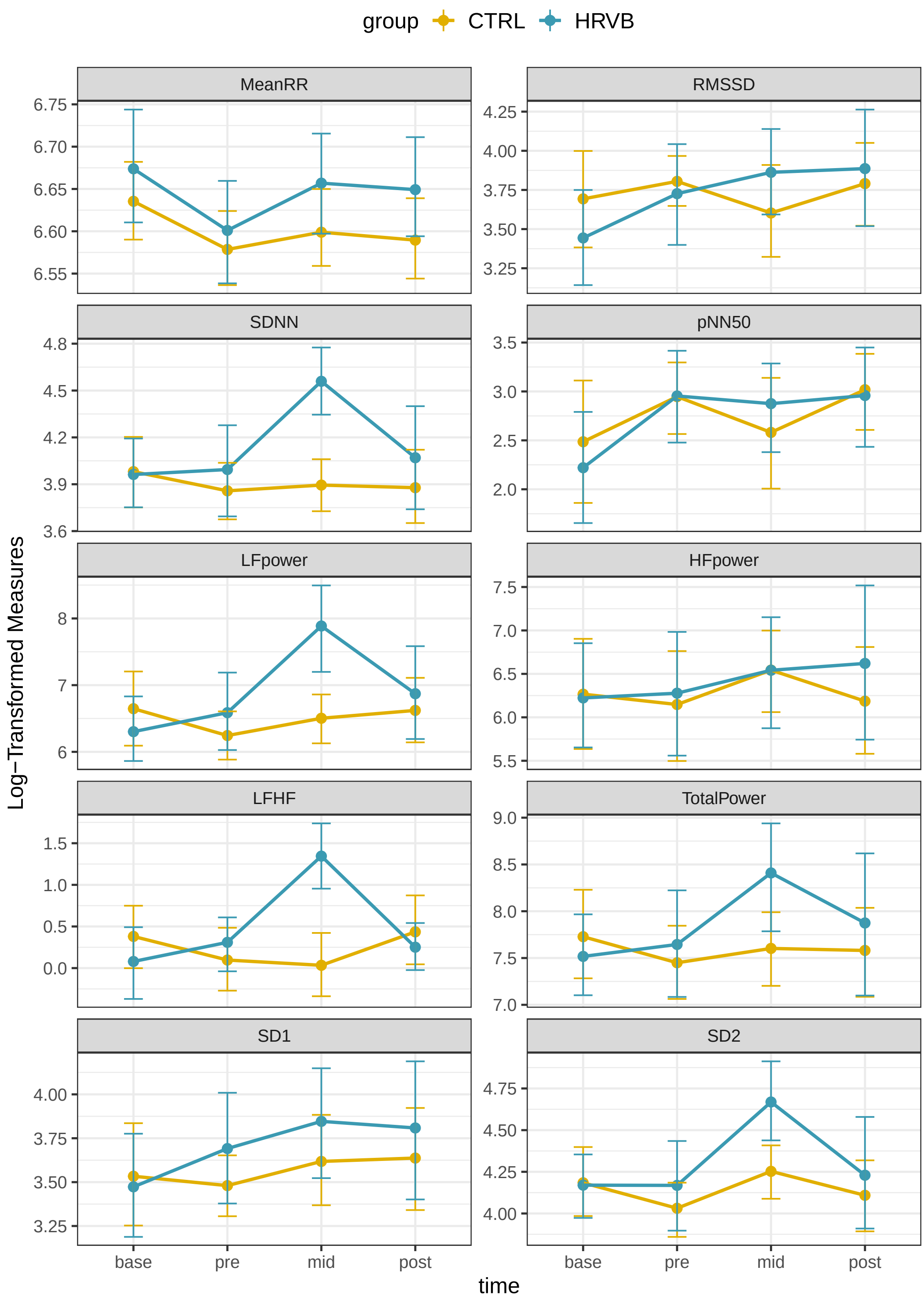


Figure 6.12: Mean of HRV Measures with 95% CI in Baseline, pre-, mid-, and post-Intervention.

- **Time-Domain Measures**

  There were statistically significant interaction effects of group$\times$time on SDNN, ($\chi^2(3) = 13.7,\ p = .003,\ \omega^2 = .09$); thus, a post-hoc analysis was conducted to break down the interaction. The results revealed a significant difference in SDNN value between the CTRL and HRVB groups mid-intervention ($b = .66,\ t(36) = 3.7,\ p = .02$, Hedges' g = 1.47), and this difference was significantly higher than the SDNN values of the HRVB group across all time points ($p < .01$ for all). On average, the post-intervention SDNN maintained a higher value in the HRVB group (M = 79.2 ms, SE = 14.6 ms) than the CTRL group (M = 55.9 ms, SE = 7.8 ms). However, this difference was not significant ($b = .19,\ t(36) = 1.6,\ p = .96$), representing a small effect size (Hedges' g = .27).

  Although there were no significant interaction effects on MeanRR ($\chi^2(3) = 5.4,\ p = .15,\ \omega^2 = .02$), time had a significant main effect on MeanRR, ($\chi^2(3) = 40.1,\ p < .0001,\ \omega^2 = .29$). The post-hoc analysis revealed a significant decrease in MeanRR pre-intervention compared to the baseline measurement for both groups, indicating HR elevation during the first cognitive stress task for the CTRL group ($b = .057,\ t(108) = 4.2,\ p < .01,\ r = .38$) and the HRVB ($b = .073,\ t(108) = 5.43,\ p < .001,\ r = .46$). Moreover, the results showed a significant increase in MeanRR for the HRVB group mid-intervention ($b = .056,\ t(108) = 4.2,\ p < .0001,\ r = .37$) compared to pre-intervention. The average MeanRR for the HRVB group significantly increased post-intervention compared to pre-intervention ($b = .05,\ t(108) = 3.6,\ p = .01,\ r = .33$). On average, the post-intervention MeanRR was higher in the HRVB group (M = 778.7 ms, SE = 24 ms) than in the CTRL group (M = 731.4 ms, SE = 18 ms). However, this difference was not significant ($b = .06,\ t(36) = 1.5,\ p = .81$), representing a small effect size (r = .24).

Group and time showed no statistically significant interaction or main effects on the RMSSD ($\chi^2(3) = 7.2,\ p = .06,\ \omega^2 = .04$) and pNN50 measures ($\chi^2(3) = 3.3,\ p = .35,\ \omega^2 = .001$). However, the interaction of RMSSD represented a small-to-medium effect size. The average RMSSD was higher for the HRVB group than the CTRL group mid-intervention (M_{diff} = 13.8 ms, SE = 11.7), and post-intervention (M_{diff} = 16 ms, SE = 16.9). However, these differences were neither significant mid-intervention ($b = .26,\ t(36) = 1.2,\ p = .93$, Hedges' g = .39) nor post-intervention ($b = .10,\ t(36) = .22,\ p = .99$, Hedges' g = .12), representing small-to-medium and small effect sizes, respectively. In addition, RMSSD for the HRVB group noticeably increased from baseline to post-intervention (M_{diff} = 27.7 ms, SE = 17.9, $b = .44,\ t(108) = 3.1,\ p = .05,\ r = .29$), representing a medium effect size.

- **Frequency-Domain Measures**

  There were statistically significant interaction effects of group×time on LF power ($\chi^2(3) = 13.8,\ p = .003,\ \omega^2 = .09$), and LF/HF ($\chi^2(3) = 32.9,\ p < .001,\ \omega^2 = .23$). The post-hoc analysis revealed that the HRVB group had a significant mid-intervention increase in LF power ($b = 1.4,\ t(36) = 3.45,\ p < .05$, Hedges' g = 1.2), and LF/HF ($b = 1.3,\ t(36) = 4.65,\ p < .001$ Hedges' g = 1.4) compared to the CTRL. Overall, these measures showed significant differences between the HRVB group mid-intervention and at all other time points ($p < .05$ for all).

  According to the descriptive statistics and exploratory graphs, the average HF power at post-intervention was greater for the HRVB group than the CTRL group (M_{diff} = 2066.8, SE = 1291.6). This difference was not significant ($b = .43,\ t(36) = .88,\ p = .98$), representing a small effect size (Hedges' g = .24). Similarly, the average total power was higher for the HRVB group than the CTRL mid-intervention (M_{diff} = 4922, SE = 1587) and CTRL post-intervention (M_{diff} = 6366, SE = 3750). However, the differences were not significant ($p > .05$), representing medium-to-large-sized (Hedges' $g = .74$) and small-sized effects (Hedges' $g = .20$) for the mid-intervention and post-intervention, respectively.

- **Non-Linear Analysis**
  Time had a significant main effect on SD2 ($\chi^2(3) = 15.56, p < .005,\ \omega^2 = .02$). Tukey's post-hoc analysis showed a significant difference in SD2 for the HRVB group at the mid-intervention time point in comparison to the other time points ($p < .05$ for all). Moreover, the descriptive statistics and exploratory graphs showed that mid-intervention SD2 was higher in the HRVB group (M = 121.97, SE = 13.7) than the CTRL (M = 74.33, SE = 5.6); however, this difference was not significant ($b = .42,\ t(36) = 2.42,\ p > .05$), representing a large-sized effect (Hedges' $g = .86$). There were no significant differences in SD1 ($\chi^2(3) = 5.4,\ p = .14,\ \omega^2 = .008$).

Regression Analysis

Based on the statistical results presented in the previous subsection, a series of regression analyses was conducted to examine whether the changes in HRV measures mid- and post-intervention could predict participants' cognitive performance and serenity scores.

First, a binomial logistic regression analysis was performed to determine whether the intervention, as characterised by SDNN and LF power, could classify participants according to their appropriate groups. Second, a linear regression analysis was performed to explore whether paced breathing, as characterised by SDNN, could predict serenity levels. Finally, a linear regression analysis was performed to predict cognitive performance post-intervention using MeanRR as a predictor variable, given that the previous analysis identified it as the HRV measure most influenced by the biofeedback practice. The RMSSD measure was also used as a predictor of both linear regression analyses due to its documented association with vagal tone and resistance to the respiratory influence compared to HF power (Laborde et al., 2017).

- **Group Classification**

  Binomial logistic regression was performed to ascertain the effect of the mid-intervention SDNN and LF power on the likelihood of correctly predicting the group to which participants were assigned. The logistic regression models were statistically significant for SDNN ($\chi^2(1) = 17.2,\ p < .0001$) and LF power ($\chi^2(1) = 21.6,\ p < .0001$). The SDNN model (B = 3.28, B SE = 1.09, z = 3.02, $p < .01$) significantly accounted for 48.5% (Nagelkerke R^2: 95% CI [4.4, 350]) of the intervention variance. Further, the LF power model (B = 2.4, B SE = .8, z = 3.0, $p < .01$) explained 57.8% (R^2: 95% CI [3.1, 77.7]) of the variance.

- **Serenity Prediction**

  After adjustment for age, the linear regression model established that SDNN mid-intervention could statistically predict serenity levels, as assessed by the difference in scores pre- and mid-intervention ($S_{\text{mid-pre}}$: $F(2, 35) = 3.08,\ p = .03$). SDNN accounted for 14.5% of the explained variance in the $S_{\text{mid-pre}}$ with an adjusted R^2 of 10.1% (B = 2.02, SE B = .86, t = 2.3, 95% CI [.28, 3.7]).

  To evaluate the association between relaxation levels and vagal tone, the post-intervention RMSSD was selected as a predictor and serenity levels were selected as an outcome, as assessed by the difference in pre- and post-intervention scores ($S_{\text{post-pre}}$). After adjustment for age, the results demonstrated that the RMSSD could not statistically significantly predict $S_{\text{post-pre}}$ ($F(2, 35) = .07,\ p = .93$) and the RMSSD (B = .21, SE B = .58, t = .35, 95% CI [-.97, 1.3]) accounted for .3% of the variance in $S_{\text{post-pre}}$.

- **Cognitive Performance Prediction**

  After adjustment for age, a linear regression model established that MeanRR post-intervention could statistically predict cognitive performance, as assessed by the difference in CR pre- and post-intervention (CR_{diff}: $F(2, 35) = 3.9,\ p = .03$). The adjusted MeanRR significantly accounted for 18.2% of the explained variation in the CR_{diff} with an adjusted R^2 of 12.3% (B = 50.5, SE B = 25.1, t = 2.01, 95% CI [11, 101]).

  To assess the neurovisceral integration model, the post-intervention RMSSD was selected for the linear regression analysis. After adjustment for age, the results demonstrated that the post-intervention RMSSD (B = .85, SE B = 4.2, t = .20, 95% CI [.05, 1.44]) accounted for 8.9% of the variance in CR_{diff}, with an adjusted R^2 of 3.7% ($F(1, 36) = 1.7,\ p = .19$).

6.6 Discussion

The hypotheses of this study were designed to gain a more in-depth understanding of the influence of a single HRVB session on short-term affective states, cognitive performance, and physiological responses. The subsequent sections provide a thorough discussion of each investigated area as well as a summary of the overall findings, followed by the limitations of the study.

6.6.1 Affective States

Several self-reported affective states were investigated in this study, including attentiveness, fatigue, mood, serenity, and stress. The attentiveness and serenity components of the PANAS-X questionnaire revealed positive results for the attention scores and relaxation levels after the HRVB intervention. These findings are in line with those reported in previous studies focused on attention control (de Bruin et al., 2016) and relaxation (Clamor et al., 2016; Lin et al., 2020; Prinsloo et al., 2013; Van Diest et al., 2014; Zaccaro et al., 2018). The present study extended these findings by examining participants' subjective perception of their own attentiveness and relaxation following the stress task and HRVB session. Regarding serenity levels, Lehrer and Gevirtz (2014) stipulated that the mechanisms underlying HRVB induce a relaxation response by stimulating parasympathetic activity mediated by vagal tone. The attentiveness score outcomes in this study suggest a link to improved performance in the cognitive task after the biofeedback intervention. During the debriefing session, one participant in the HRVB group commented: "The deep breathing practice helped me think clearly about strategies to solve the cognitive task".

Although there was no statistical evidence regarding the influence of paced breathing on perceived fatigue, the HRVB group reported slightly higher average scores in comparison to the CTRL immediately after the intervention. This finding could be due to the participants' lack of familiarity with paced breathing exercises, which resulted in a dyspnoeic or uncomfortable experience. In the same vein, You et al. (2021b) explained that the increase in perceived stress following a 5-min paced breathing exercise in their study was due to breathing discomfort, which is typical for individuals who are unfamiliar with the practice. Moreover, there were no significant changes in the reported stress and mood scores, which could be attributed to the unidimensionality aspect of using single-item Likert scales to assess complex constructs, such as stress and mood (Hasson & Arnetz, 2005).

6.6.2 Cognitive Performance

For the cognitive performance aspect of this study, the HRVB group performed better than the CTRL in the second N-back task, which assessed participants' working memory capacity with respect to CR. These results are consistent with

Prinsloo et al.'s (2011) findings regarding improvement in cognitive performance (i.e., inhibitory control measured using a Stroop task) after a single HRVB session. This observed significant increase in CR could be attributed to the HRVB intervention, which stimulated the vagus nerve. In particular, previous studies have linked the activation of parasympathetic activity with working memory and attention-based tasks (Forte et al., 2019; Hansen et al., 2004; Hansen et al., 2003).

However, there was no significant difference in RT between the two groups post-intervention in the present study, which is in direct contrast to Prinsloo et al. (2011). This rather contradictory result may be due to the experimental protocol as the previous study advised participants to consider speed when responding, whereas participants were not similarly advised in this study. Another possible explanation may be that this study looked at RT for CR to accurately quantify processing speed (Ratcliff, 1993). Mahinrad et al. (2016) found that poor processing speed and long RT in cognitive functioning evaluated by a Stroop task were associated with low HRV measures. However, the authors analysed HRV signals using a 10-s segment, while the present study analysed HRV signals using a 5-min segment. There is a well-established trade-off between accuracy and response time in cognitive activities: individuals compromise accuracy for speed, or vice versa (Donkin et al., 2014; Franzon & Hugdahl, 1987; Wylie et al., 2009). Further, Mahinrad et al. (2016) focused on a specific age group (i.e., older participants), thus limiting the generalisability of their findings to younger age groups.

6.6.3 Physiological Measures

Similar to the study discussed in Chapter 5, this study examined the impact of cognitive stress and paced breathing on HRV measures. However, the stress task in this study was specifically focused on working memory. Further, the paced breathing exercise utilised a prolonged exhalation period. Longer exhalations have been shown to be a stimulating protocol for notable improvements in cardiac vagal tone (Van Diest et al., 2014), as indexed by RMSSD measure (Laborde et al., 2021). Overall, the results of the present study were similar to those of the previous study. In addition, they are consistent with earlier research in which LF/HF, LF power, and SDNN increased during the paced breathing exercise and MeanRR decreased during the stress task (see Section 5.6).

An interesting finding from the current study is that the MeanRR of the HRVB group maintained a high value during the cognitive task post-intervention compared to pre-intervention, suggesting that a single HRVB session can have short-term lasting effects on HRV after returning to a normal rate of breathing. However, it is possible that this finding may be biased due to the uncertainty of the relationship between MeanRR and vagal tone (Shaffer & Ginsberg, 2017).

Contrary to the results reported in You et al. (2021a), the present study did not find significant group differences in RMSSD during the intervention; however, these results are in line with those of Laborde et al. (2019b). While this study implemented a paced breathing rate similar to that in You et al. (2021a), the discrepancy in results could be due to the number and duration of paced breathing exercises (e.g., three 5-min sessions), type of participant (e.g., athletes), or type of control (e.g., watching TV).

The BP measurements were collected at baseline and after the two cognitive tasks. After the first cognitive task (pre-intervention), SBP increased for both groups relative to baseline. However, there were no significant changes in the DBP measurements. This finding corroborates earlier research concerning the influence of stress on SBP elevation (Hjortskov et al., 2004; Steffen et al., 2017). Conversely, some studies have reported changes in SBP but not DBP (Vrijkotte et al., 2000). A possible explanation for the relative insensitivity of DBP reactivity in this study is the mild effect elicited by the chosen stress inducer. In comparison to the CTRL, the HRVB group exhibited lower SBP levels (i.e., similar to baseline) in response to the second cognitive task post-intervention. This finding is seemingly consistent with the results of Steffen et al. (2017), which showed that participants who performed resonance breathing had lower SBP reactivity in response to the stress task than participants in the other study groups. The greater effects of paced breathing on SBP than DBP could be explained by the direct relationship between SBP and the baroreflex activity (Lehrer et al., 2020).

6.6.4 Overall Discussion

Taken together, the findings partially support H1 because the HRVB group reported higher attentiveness and serenity scores post-intervention compared to the CTRL. In contrast, there was insufficient evidence to claim improvements in stress, mood, and fatigue following the HRVB intervention. In addition, the findings partially support H2 because the HRVB group performed better in the cognitive task compared to the CTRL, as assessed by CR. However, no differences were found with respect to RT.

At the physiological level, the findings partially support H3 because the HRVB group exhibited lower SBP in response to the post-intervention stress task than the CTRL, while no significant effect was observed on DBP. Further, the SDNN and frequency-domain measures significantly increased during the intervention, whereas MeanRR increased for the HRVB group post-intervention. Further, higher levels of SDNN during the intervention predicted improved relaxation, as reported by the serenity questionnaire. In addition, the findings support H4 because increased MeanRR post-intervention predicted high cognitive performance, as assessed by CR. Moreover, increased SDNN mid-intervention predicted relaxation levels, as indexed by the serenity scores.

Although vagal tone, as reflected by RMSSD, was higher in the HRVB group post-intervention compared to baseline with medium effect size, the results were not statistically significant in comparison to the CTRL. Therefore, there was a lack of significant association between vagal tone and improved cognitive performance and relaxation levels. These findings are in agreement with prior research demonstrating that a single-paced breathing session does not sufficiently improve RMSSD after the session (Laborde et al., 2019b; You et al., 2021a). Consequently, the present study obtained a similar conclusion regarding post-intervention RMSSD and vagal tone, despite the previous two studies not including a biofeedback component in their design.

6.7 Limitations

There are a number of limitations in the present study related to the biofeedback protocol. First, the lack of participants' familiarity with paced breathing exercises may have posed challenges in correctly performing the activity. Although the exercise duration was intentionally selected to be short (6 min) to minimise discomfort in participants unfamiliar with the exercise, a better strategy may be to adopt multiple consecutive short sessions with breaks in between, as in You et al. (2021a). Second, all participants performed the breathing exercise at the same rate of 6 breaths/min rather than determining the RF for each participant. Although several studies have indicated similar physiological behaviour with 6 breaths/min during the exercise, Steffen et al. (2017) observed differences in self-reported mood and SBP between breathing at RF and RF+1. Consequently, future studies could investigate the distinctions between RF and breathing at a fixed rate after the paced breathing exercise at a psychological level.

6.8 Chapter Summary

This chapter addresses SRQ3 by presenting the results of an RCT experiment design used to investigate the short-term effects of HRVB on affective states, cognitive performance, and physiological responses. The experiment involved two independent groups: a paced breathing intervention group (HRVB) and a spontaneous breathing group (CTRL). Affective states (i.e., attentiveness, fatigue, mood, serenity, and stress) were assessed using self-reported questionnaires. Cognitive performance was assessed using cognitive performance in a working memory task based on computer-logged CR and RT data. The physiological data involved HRV and BP measures.

Overall, the findings prominently indicate improvements in relaxation levels and cognitive performance following the HRVB intervention, as reflected by the serenity and CR scores, respectively. Moreover, the HRVB group exhibited lower SBP reactivity to the stress task compared to the CTRL. Although HRV measures were not statistically significant post-intervention (except for MeanRR), they were on average higher in the HRVB group compared to the CTRL, with small-sized effects. In addition, regression analyses established that SDNN mid-intervention and MeanRR post-intervention could predict relaxation levels and cognitive performance, respectively.

Having demonstrated the influence of HRVB on various physiological and psychological responses, the next chapter focuses on the employment of ML algorithms to identify stress and relaxation levels from the examined HRV measures.

CHAPTER 7

Stress Recognition

This chapter leverages the datasets and main findings of the previous two chapters to develop robust predictive models for supervised learning algorithms capable of recognising stress and relaxation states using short-term and ultra-short-term heart rate variability measures.

7.1 Overview

Recognising various levels of stress can assist in the development of strategies for early intervention, stress management, and risk prevention to promote mental health and well-being (Hazer-Rau et al., 2020). A growing number of researchers have investigated stress detection by developing predictive models using ML algorithms based on physiological data (Bobade & Vani, 2020; Castaldo et al., 2019; Dalmeida & Masala, 2021; Sarkar & Etemad, 2020; Theeng Tamang et al., 2020). Among the various physiological measures studied, HRV is recognised as a significant biomarker for monitoring mental stress responses by reflecting the activity levels of the parasympathetic and sympathetic branches of the ANS.

Although limited datasets are commonly used in affective computing and psychophysiological research, caution must be applied in the development of ML algorithms to avoid biased conclusions regarding model performance. Schmidt et al.'s (2019) review of affect recognition using ML found that most of the studies included data collected from fewer than 40 participants (43 out of

46 studies), and only one study included more than 100 participants. Moreover, the review indicated a wide gap in the reported performance metrics among the 46 studies, with the accuracy rates ranging from 40% to 97%. Concerns were raised in the domains focused on biomedical studies (Foster et al., 2014) and psychiatric disorders (Cearns et al., 2019) as significant variation in ML model accuracy with the use of limited datasets could indicate performance overestimation or methodological issues.

In the present study, predictive models were developed to classify stress levels based on HRV measures, as addressed by the following research sub-question:

> **SRQ4:** How can robust supervised learning algorithms recognise stress and relaxed states for eventual deployment in real-time systems?

Particular attention was given to overcoming the primary shortcomings related to ML algorithm implementation and interpretation identified in the literature. These shortcomings included data segmentation, feature selection, and model evaluation, which are associated with the use of limited datasets (see Section 7.2). Accordingly, this study was designed in accordance with the recommendations for ML with small datasets (Cearns et al., 2019; Foster et al., 2014; Stevens et al., 2020; Vabalas et al., 2019), and these recommendations are summarised in Section 7.2.2.

7.2 Related Work

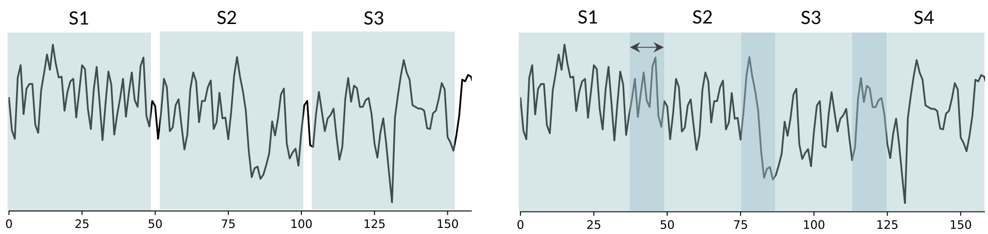
7.2.1 Methodological Limitations

Stress detection using ML techniques has gained considerable attention in the fields of affective computing and psychophysiology (Dalmeida & Masala, 2021; Gedam & Paul, 2021; Healey & Picard, 2005). Recent advancements in technology, particularly wearable devices, have provided non-invasive approaches for physiological data collection and analysis. However, a reliance on small sample sizes in conjunction with certain common methodological limitations pose problems for the implementation and interpretation of ML algorithms. These problems include overfitting, overly optimistic performance, and generalisability issues.

Data Segmentation

The principal issue related to data segmentation concerns the tendency to increase the dataset size by dividing the physiological signals from each participant into multifold segments, which violates the statistical assumption of the independence of observations (Chen et al., 2021; Dalmeida & Masala, 2021; Oskooei et al., 2021). Hence, the resulting segments are dependent because data were generated from the same participant, which can cause data leakage by way of dependent observations in the training and testing sets (Foster et al., 2014).

An additional source of potential dependency is the adoption of an overlapping window segmentation approach instead of a non-overlapping approach (Schmidt et al., 2018; Smets et al., 2016; Tervonen et al., 2021). In the overlapping approach, the observations are not only generated from the same participant but the physiological data are also partially shared among the observations (see Figure 7.1).



(a) Non-Overlapping Segmentation (b) Overlapping Segmentation

Figure 7.1: Physiological Data Segmentation Approaches with a 50-s Window Size.

To detect the severity of panic attacks in 10 participants, Rubin et al. (2016) used an overlapping window approach with HRV data to generate a vast number of observations, reaching a maximum of 1,700 samples and, by extension, boosting the size of the training and testing sets. Similarly, Schmidt et al. (2018) adopted an overlapping window approach with a shift of .25 s for physiological data collected from 15 participants. However, they employed a subject-independent validation strategy to mitigate data leakage (see Section 7.2.1). In a fear classification study, Petrescu et al. (2021) used overlapping and non-overlapping segmentation techniques on a dataset consisting of 32 participants. They reported equivocal results regarding the ML model performance for each segmentation approach. However, it is not clear to what extent classification accuracy is impacted by the use of an overlapping technique vs. a non-overlapping one (Anusha et al., 2018). In fact, Dehghani et al. (2019) demonstrated that improved model performance is associated with the use of dependent observations and the employment of an inadequate validation strategy.

Data leakage provides overly optimistic estimates of generalisation performance because dependent observations are presented in the training and testing sets. Further details regarding the theoretical and mathematical derivations of performance overestimation are discussed in Cawley and Talbot (2010) and Hastie et al. (2009, p. 228). In Castaldo et al.'s (2019) remarkable study, mental stress levels were classified by mitigating the violation of the independence assumption using two considerations. First, the authors did not use any segmentation methods in their implementation. Second, the ML models were trained and tested on different groups of participants using data collected in the same experiment and analysed using a subject-independent validation approach (see Section 7.2.1). They reported high classification performance for the selected ML models, with a minimum accuracy rate of 88%. However, the generalisability of their findings is limited due to the extremely small training (25 participants) and testing sets (17 participants).

Model Evaluation and Selection

Given the general limitations of small sample sizes, the strategy employed for model evaluation can aggravate the interpretation of ML performance. Several validation strategies are commonly used in the implementation of supervised ML algorithms, such as the hold-out method and cross validation (CV) techniques (Hastie et al., 2009). The latter is more often employed in the context of limited datasets because of its ability to utilise the entire dataset in model fitting and evaluation.

K-fold is a prominent CV technique that randomly splits the dataset into K groups and then trains the model iteratively on the K-1 groups while keeping the remaining group for validation (Hastie et al., 2009). Subsequently, overall performance is calculated as the average accuracy rate resulting from all K trials. However, random splitting with dependent observations poses a data leakage problem as the training and testing sets may include data segments from the same participant. As briefly discussed in the previous section, data leakage leads to biased and overly optimistic generalisation performance estimates. Recent research has suggested splitting the data per participant using a subject-independent CV, also known as a leave-one-out CV, to limit the effect of the dependent observations on the development and evaluation of the ML models (Dehghani et al., 2019; Esterman et al., 2010). The leave-one-out CV is an example of the K-fold method, where K is the total number of observations or participants. In a review of affect recognition conducted by Schmidt et al. (2019), 13 studies (out of 46) used the K-fold CV, while the remaining studies incorporated variations of the leave-one-out CV. This demonstrates that the leave-one-out CV is the preferred approach to mitigate the violation of the independence assumption within the context of affective computing applications.

Hyperparameter selection is commonly performed prior to model evaluation, although the use of a standard CV procedure with both processes can cause model selection bias. In particular, the use of the same test set in each process can introduce overly optimistic estimates of the expected generalisation performance (Cawley & Talbot, 2010). Consequently, the nested CV technique can be used to manage both model evaluation and hyperparameter selection as integral processes, albeit with different validation/test sets. Further details about the nested CV are discussed in Section 7.4.4.

Feature Selection

An additional issue relates to the number and choice of features employed in the ML classifiers. The current literature has reported two extremes in the feature selection process: the inclusion of all collected physiological measures regardless of dataset size and the inclusion of features irrelevant to the investigated problem (e.g., behavioural, clinical; Cho et al., 2017; Coutts et al., 2020; Schmidt et al., 2018).

Vabalas et al. (2019) delineated the increased likelihood of overfitting when many features were included relative to the sample size, particularly with small datasets. Overfitting is often caused by complicated model development that results in memorisation of the training data rather than learning of the underlying patterns to be applied in future predictions (Hawkins, 2004). As a result, the ML model performs poorly on the testing set, despite high performance in the training set. Therefore, a large number of features and redundant features increase model complexity, leading to inaccurate ML performance assessment (Ying, 2019).

Irrelevant feature selection based on the context of the examined clinical condition also restricts contextual interpretation of the predictive models. For instance, Castaldo et al. (2016) selected non-linear HRV measures to classify stress levels based on a statistical correlation analysis between the features and target. However, the physiological rationale behind feature selection was not discussed. Additionally, an analysis of 30-s segments was performed to obtain VLF power (.0033-.04 Hz) from the HRV frequency domain as an ML feature (Dalmeida & Masala, 2021). However, a segment with a minimum length of 5 min was found to be necessary for the robust computation of frequency components in the VLF band (Shaffer & Ginsberg, 2017). From a clinical perspective, a physiological measure calculated from an ultra-short segment may increase the risk of interpreting results that are not representative of their actual medical meanings because the number of samples is insufficient for reliable analysis (Malik et al., 1996). Thus, a reliability analysis of UST measures is essential prior to feature selection.

7.2.2 Recommendations

Given the limitations discussed in the previous section, the following subsections provide recommendations to deal with limited datasets and avoid overfitting and performance overestimation, as advocated by Cearns et al. (2019), Foster et al. (2014), Stevens et al. (2020), and Vabalas et al. (2019).

Validation Strategy

It is well established that the larger the dataset, the better the ML performance. However, independence among the observations should be considered when dealing with data generated from the same participant or obtained from segmentation, particularly when splitting the dataset into training and testing sets to avoid data leakage during model selection. To reduce the effect of the dependent observations, an appropriate validation strategy should be implemented. The leave-one-out CV technique is particularly effective for small datasets with dependent observations (Foster et al., 2014).

Moreover, performance overestimation, especially with small datasets, may arise during model selection from using the same validation/test set in the hyperparameter selection and model evaluation processes. In the Scikit-Learn Python package, model training and evaluation for hyperparameter selection are implicitly conducted by automatically refitting the model using GridSearchCV (Buitinck et al., 2013; Scikit-Learn Developers, 2022). Hence, the nested CV approach is proposed as a mitigation strategy for selection bias and performance overestimation (Cawley & Talbot, 2010; Cearns et al., 2019; Vabalas et al., 2019).

Feature Optimality

Features should be rationally selected based on the clinical or physiological motivation of the investigated ML problem to facilitate the contextual interpretation of algorithm performance (Remeseiro & Bolon-Canedo, 2019). After determining the most relevant features, several techniques can be used to select the optimal features, such as correlational analysis or feature elimination methods. Moreover, principal component analysis can be used to reduce high dimensionality (Chandrashekar & Sahin, 2014).

To minimise the effect of performance overestimation and reduce computational costs, the selected features should be limited to a reasonable feature-to-sample ratio (Vabalas et al., 2019). A common practice in biomedical research using small datasets is to choose one feature for every 10 independent observations (Foster et al., 2014).

7.3 Proposed Classification Approach

To achieve results comparable with the literature reviewed in this chapter, six common supervised ML algorithms were selected: logistic regression (LR), decision trees (DT), k-nearest neighbours (KNN), Naive Bayes (NB), random forest classifier (RFC), and support vector machine (SVM). These models were fitted on the datasets described in Chapters 5 and 6. The nested CV method was used to perform hyperparameter selection and model evaluation as integral processes using the leave-one-group-out (LOGO) CV, which is a variation of the leave-one-out method (Maleki et al., 2020). Further details about the LOGO CV are described in Section 7.4.4.

As discussed in Chapters 5 and 6, the HRV data of each participant in the dataset were assigned three labels based on the condition of data acquisition: 1)

neutral (baseline), 2) **stress** (TSST or N-back stress task), and 3) **relaxed** (paced breathing exercise). In a preliminary analysis of a three-class ML classifier using DT, the algorithm showed high accuracy rates in identifying the neutral (90%) and relaxed states (97%) but failed to distinguish the stress from neutral states (34%). This confusion between the neutral and stress states could be due to the moderate effect of mental stressors on HRV measures, as discussed in Sections 5.6 and 6.6. Therefore, two independent binary classifiers were implemented to differentiate the stress state from each non-stress state: 1) stress vs. neutral and 2) stress vs. relaxed.

To assess generalisability, the ML model that showed the best performance resulting from the nested CV method was evaluated using two combined independent datasets with 300-s short-term (ST)[1] HRV recordings. Moreover, one 60-s UST segment was extracted from the middle of each HRV recording to assess classification performance for potential future deployment in wearable devices. The ML algorithms were implemented using the Scikit-Learn Python package (Barupal & Fiehn, 2019). Figure 7.2 illustrates the overall process, including data preprocessing, feature selection, model selection, and evaluation.

[1]The abbreviation "ST" is only used in this chapter.

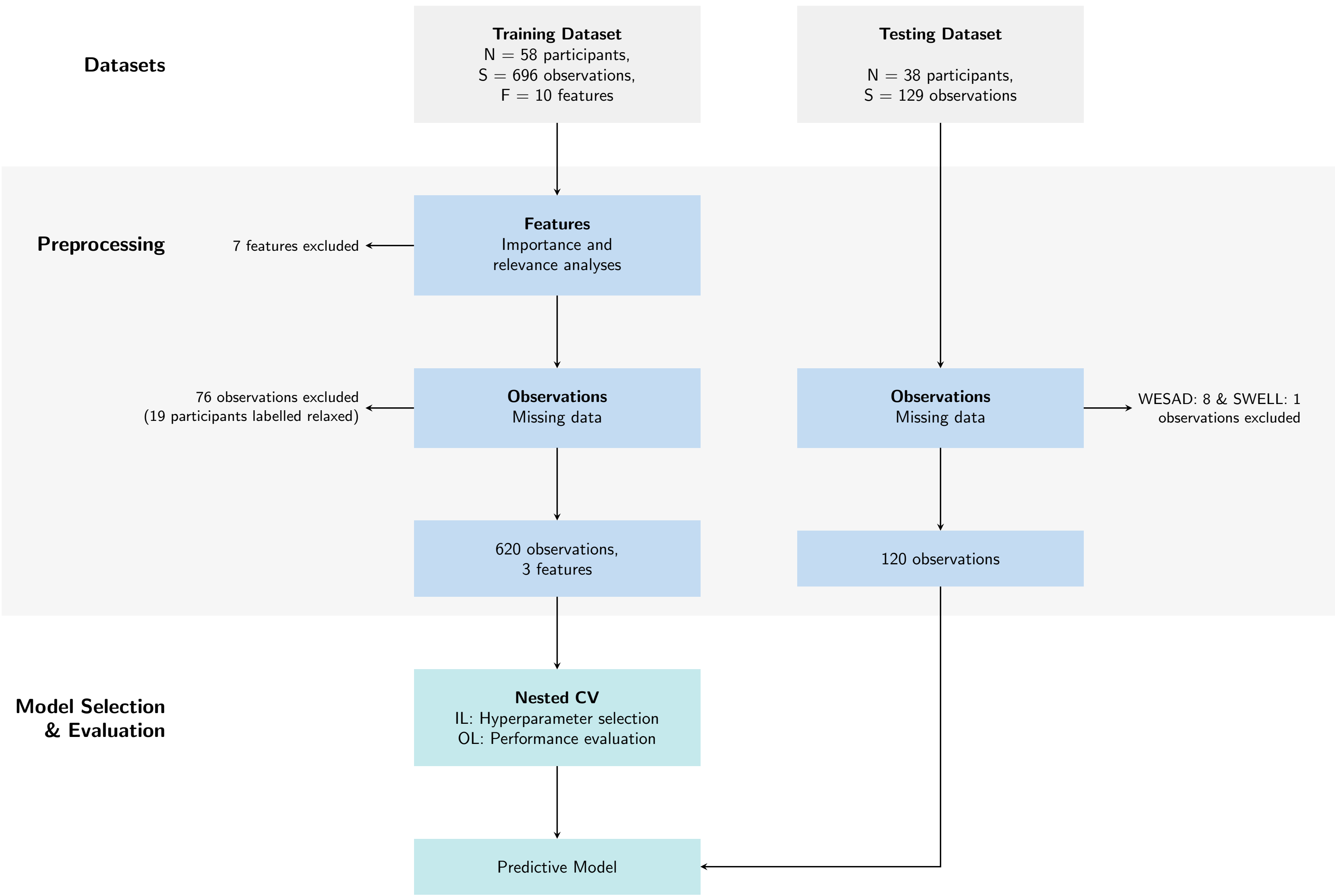


Figure 7.2: A flowchart of the ML process including dataset split, preprocessing, model selection and evaluation.

Note. IL: Inner Loop, OL: Outer Loop. Adapted from Stevens et al., 2020

7.4 Methods

7.4.1 Dataset

As described in the UST exploratory (20 participants; Chapter 5) and biofeedback studies (38 participants; Chapter 6), the collected HRV data exhibited similar HRV trends in response to the employed stress tasks and paced breathing exercise. Therefore, these two datasets were combined to develop ML binary stress classifiers. In total, the dataset included 58 participants, and three labelled HRV recordings were applied to each under neutral, stress, and relaxed conditions. This dataset (i.e., original dataset) was used as the training set for the ML model development.

As the data for the second stress task in the biofeedback dataset were collected after paced breathing, they were discarded to obtain a consistent protocol. Additionally, because participants in the CTRL group (19) did not perform the paced breathing exercise, the data on their relaxed state were excluded. Each recording was divided into four non-overlapping segments with a duration of 60 s. Hence, the total number of observations was 620, of which 232 samples were labelled neutral (58 participants $\times$ 4 segments), 232 samples were labelled stress (58 participants $\times$ 4 segments), and 156 were labelled relaxed (39 participants $\times$ 4 segments). Further details about the participants in the exploratory and biofeedback studies are provided in Sections 5.4 and 6.4, respectively.

The generalisability of the developed ML models was assessed using two external datasets: WESAD (15 participants; Schmidt et al., 2018) and SWELL (23 participants; Koldijk et al., 2014). Several publicly available datasets were reviewed, but these two were selected because they included data on mental stress tasks, 5-min or longer HRV recordings, and detailed explanations of the experimental methods and protocol (see Table B.1). The data labelled stress and relaxed for eight participants were excluded from the WESAD dataset because they performed the paced breathing exercise before the stress task. As the present study was focused on three states (i.e., neutral, stress, and relaxed), the HRV data collected during the amusement condition were also excluded (see Section 3.3.5). This dataset (i.e., independent dataset) was used as the testing set for the ML model generalisability assessment.

All data were filtered and checked for signal quality, resulting in the exclusion of one HRV recording in the relaxed state from the SWELL's dataset because the number of samples was insufficient for HRV analysis. Therefore, the total number of observations was 120: 38 samples were labelled neutral, 53 were labelled stress (23×2 SWELL + 7 WESAD), and 29 were labelled relaxed (15×2 - 1 WESAD). Further details about the participants and experimental protocol applied to each dataset are presented in Section 3.3.5.

7.4.2 Data Preprocessing

Due to the physiological differences among participants across the four datasets, all recordings were normalised based on the average HRV of each participant's baseline measurement (see Equation 7.1; Sacha, 2013; Sacha & Pluta, 2008).

$$RR(i) = \frac{RR(i)}{mean(RR_{baseline})}, \quad i = 1, 2, ..N \tag{7.1}$$

Based on the findings obtained from the UST analysis described in Chapter 5, a non-overlapping segmentation approach was used on the original dataset to divide the 300-s HRV recording into shorter segments with a window size of 60 s. To minimise dependency among segments, a 10-s gap was applied in the segmentation process, resulting in four segments per condition per participant (see Figure 7.3).

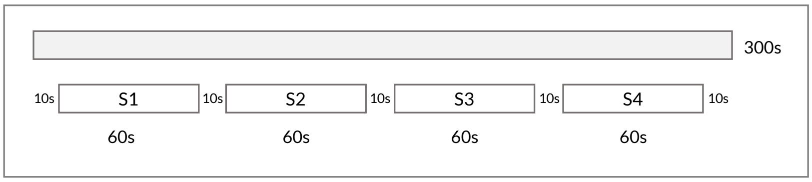


Figure 7.3: Non-Overlapping Segmentation Approach

To maintain consistency between the original and independent datasets, the ECG signals from the WESAD (700 Hz) and SWELL (2048 Hz) datasets were downsampled to 500 Hz. Subsequently, peaks were detected to extract the RR intervals using the NeuroKit2 Python package (Makowski et al., 2021). Thereafter, a 300-s segment was extracted from the centre of each HRV recording. The HRV signals were then normalised based on Equation 7.1, filtered using the adaptive threshold detection and moving average correction algorithms (see Chapter 4), and analysed using the pyHRV and Systole Python packages (Gomes, 2018; Legrand & Allen, 2022).

7.4.3 Feature Selection

This study sought to distinguish between stress and non-stress states (i.e., neutral and relaxed). Hence, different features were selected based on the purpose of the developed ML binary classifier, albeit using a similar feature selection strategy.

According to Vabalas et al. (2019), the feature-to-sample ratio in limited datasets should be reasonably low. A common practice in biomedical research using small datasets is to select one feature for every 10 independent observations (Foster et al., 2014). Thus, a maximum number of three features was selected as the original dataset consisted of 58 participants, while the two states (stress vs. non-stress) were obtained from the same participant. These criteria resulted in 29 independent observations.

The statistical analysis of the influence of stress and paced breathing on the HRV measures showed that MeanRR significantly changed from neutral to stress and from stress to paced breathing (see Section 5.5.4). Hence, MeanRR was selected as the first feature in the implementation of both ML binary classifiers because it reflected the average variation in HRV and could be reliably measured in 60-s HRV segments. SDNN was selected as the second stress vs. relaxed feature given that a significant statistical change was noted in SDNN between both states due to its association with paced breathing (Shaffer & Ginsberg, 2017). Further, it could be calculated from the 60-s segment (see Table 5.4).

To determine the significance of the remaining features in relation to the ML models, relative feature importance was calculated using an RFC implemented via Scikit-Learn, which computed a weighted average score based on the degree to which the feature reduced impurity in the tree node. Based on the results of the feature importance calculation, RMSSD and HF power were selected as the stress vs. neutral classifiers. Both measures reflected cardiac vagal tone and could be calculated from the 60-s segment (see Table 5.4; Shaffer & Ginsberg, 2017). In contrast, SD2 was selected as the stress vs. relaxed feature due to its association with LF power and potential to be computed using 60-s segments in both states (see Table 5.4; Shaffer & Ginsberg, 2017).

A summary of the importance scores of the selected features is outlined in Table 7.1. The Spearman's rank-order correlation revealed non-significant correlation coefficients among the selected features ($p > .05$). As the features had different scales, a standardisation approach was applied to numerical features by removing the mean value and dividing it by the SD, resulting in a distribution with unit variance.

Table 7.1: Feature Importance Scores

Feature	Score
Stress vs. neutral	
MeanRR	40.1%
RMSSD	30.1%
HF power	29.8%
Stress vs. relax	
MeanRR	32.5%
SDNN	34.2%
SD2	33.3%

7.4.4 Nested Cross-Validation

Model selection was performed using the CV method and divided into two main steps: hyperparameter selection and performance evaluation. These steps are often assessed using the same validation/test set. An extension of the CV method is the nested CV approach, which is performed to overcome the biased performance estimates introduced by a standard CV (see Section 7.2.2). The nested CV approach consists of two CV procedures, each of which is performed in a different loop: the inner loop is used for the hyperparameter selection, while the outer loop is used for the performance evaluation.

A specific CV method can be selected for each loop from a pool of available methods (e.g., K-fold, leave-one-out). As previously discussed, the leave-one-out method is recommended for limited datasets and dependent observations. In this study, the LOGO method was adopted to group segments associated with each participant based on participant ID (Maleki et al., 2020). LOGO is similar to leave-one-out, but it allows for the assignment of multiple observations to a single group. The total number of splits was equal to the total number of participants in the original dataset (58), which corresponds to a 58-K-fold CV procedure.

Figure 7.4 shows the overall nested LOGO CV approach process using a simplified example of four participants, with each having four associated segments. First, the segments were grouped based on participant ID. Second, the original dataset was split into N outer training validation sets, where N was the number of participants (N = 4).

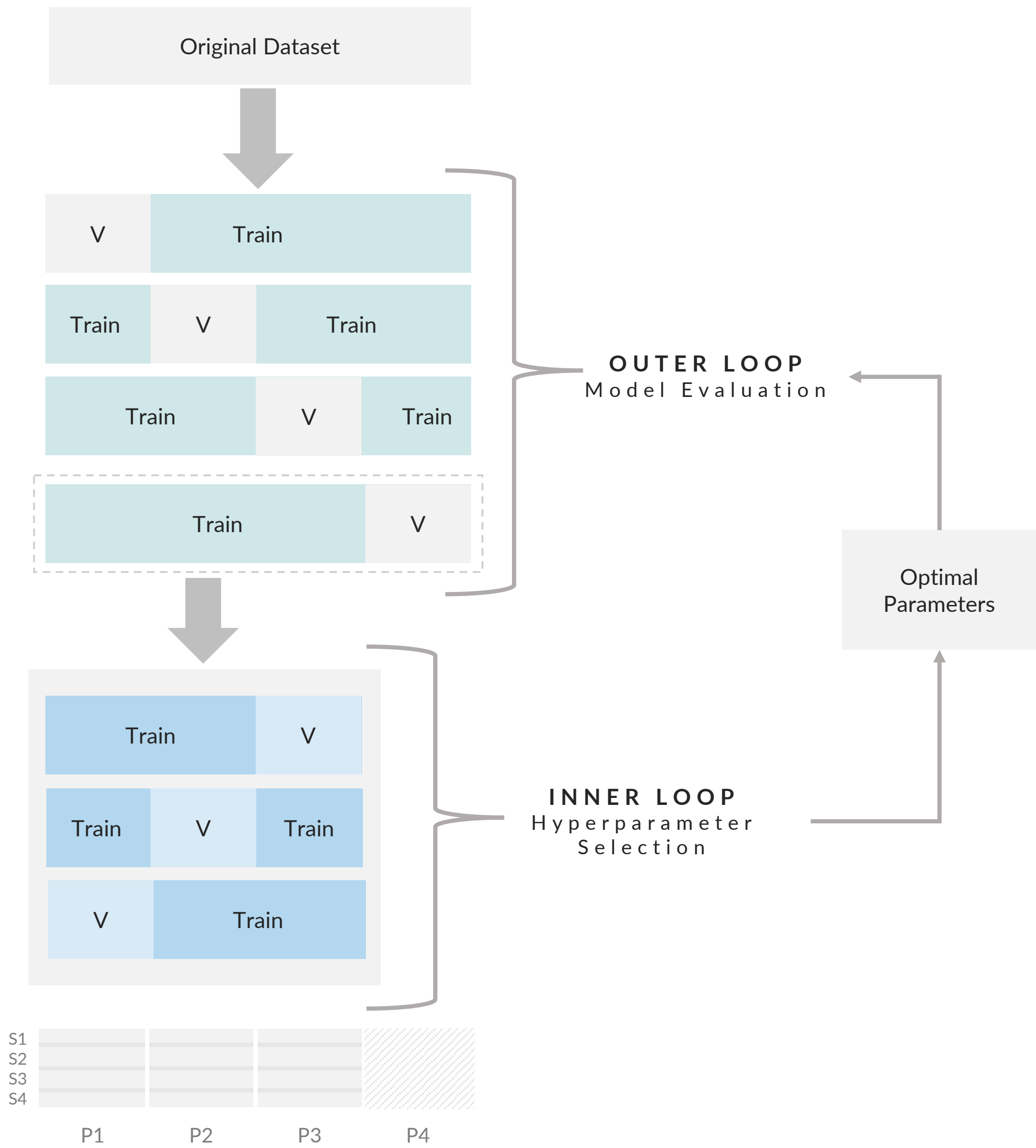


Figure 7.4: A Conceptual Illustration of the Nested CV Procedure with Four Participants, Each with Four Segments.
Note. V refers to validation set, S refers to segment number, and P refers to participant ID.

In the outer loop, a training set was selected from each iteration and inserted into the inner loop for the hyperparameter selection. In the inner loop, the selected training set was split into three (N-1) internal training validation sets. A GridSearchCV with a predefined search space for each ML algorithm was implemented to find the optimal hyperparameters, as shown in Table 7.2. The optimal hyperparameters were then selected to fit the model on the outer training set and evaluate it on the outer validation set. Four performance estimates were generated from the outer loop, and the average performance and stability were calculated for each ML algorithm. Lastly, the original dataset was retrained on the model with the highest performance and stability.

Table 7.2: Predefined Hyperparameters for the GridSearchCV

Algorithm	Hyperparameter	Value
LR	`C` (regularisation strength)	10^i, i=[-4, 4]
DT	`max_depth` `min_samples_leaf`	[1, 2, 3, 4] [.02, .04, .06, .08]
KNN	`n_neighbours`	[2, 3,.., 9]
NB	`var_smoothing`	10^i i = [-9, 0]
RFC	`max_depth` `min_samples_leaf`	[2, 3] + None [.05, .1]
SVM	`C` (regularisation strength) `kernel`	10^i, i=[-4, 4] Radial-basis function (rbf)

Although the nested LOGO CV method can be computationally expensive, it was preferentially selected over the remaining CV strategies because of its capacity to obtain unbiased true error rate estimates, thus minimising the risk of obtaining overly optimistic performance metrics (see Section 7.6).

7.4.5 Performance Metrics

ML performance was evaluated using the following metrics: accuracy, precision, recall, F1 score, confusion matrix, area under the curve (AUC), and Matthew's correlation coefficient (MCC; see Appendix F for the performance metrics equations).

- **Accuracy** measures the ratio of correct instances classified as stress and non-stress to the total predictions made by the classifier.
- **Precision** measures the ratio of correct instances classified as stress to the total predictions in the stress class. A precision of 100% indicates that the ML model generates no false positives.
- **Recall** measures the ratio of correct instances classified as stress to the actual total number of stress instances in the dataset; this metric is similar to the sensitivity measure discussed in Chapter 4. A recall of 100% indicates that the ML model generates no false negatives.
- **F1 score** represents the harmonic mean of precision and recall. It can be a useful measure when precision and recall are equally important in a classification model assessment.
- **Confusion matrix** summarises the number of correct and incorrect predictions for a binary classifier in a 2×2 table by representing the true and predicted instances in each class, thereby conveying information about the rate of observations that are correctly classified as non-stress (true negative) and stress (true positive). Moreover, a confusion matrix can determine the rate of observations that are incorrectly classified as positive or stress (false positive) and as negative or non-stress (false negative).

- **MCC** evaluates the quality of binary classifiers by summarising the confusion matrix into a single quantification measure (Baldi et al., 2000; Gorodkin, 2004; Matthews, 1975). It computes a value between -1 and 1, where 1 indicates perfect class prediction and -1 indicates inverse class prediction.

- **AUC score** is a quantification measure for the area under the receiver operating characteristic (ROC) curve. The ROC curve demonstrates the trade-off between true positive and false positive rates at different probability thresholds. Thus, AUC can be calculated as a representative measure of ROC to better interpret and assess classifier performance.

In this study, the classifications of both states (stress and non-stress) were equally important. In other words, the primary aim of the model assessment was to minimise the impact of both error rates: false positives and false negatives. Hence, the F1 score was selected as the most appropriate performance metric to provide a single representative measure for precision and recall. Moreover, the confusion matrix, accuracy, precision, recall, AUC score and MCC were employed as auxiliary performance metrics.

7.5 Results

To compare the results of this study with those of the literature, six common supervised ML algorithm models were assessed: LR, DT, KNN, RFC, NB, and SVM with a non-linear kernel (i.e., radial-basis function). The model with the best performance and stability was selected for the independent test evaluation. The stability was measured using the SD of the outer CV performance evaluation.

7.5.1 Classification of Stress and Neutral States

Model Selection

Table 7.3 shows the average performance metrics resulting from the nested CV approach used to classify stress and neutral instances from the original dataset.

Overall, the ML models had relatively low performance in classifying stress and neutral states (accuracy: 49%-60%). More specifically, the precision scores obtained by all models were less than 62% indicating a high misclassification rate of the neutral instances (i.e., high false positives). However, the recall of DT and RFC was greater than 80% indicating good performance in identifying stress instances. The AUC score of all classifiers was in the range of 60%-76%. Among all the classifiers, RFC showed the best performance and highest stability, with an F1 score of 67.3% (SD = 6%) and an MCC of 56.7%. The remaining classifiers had F1 scores in the range of 41%-59%. Hence, the RFC with the following hyperparameters was selected for the independent dataset evaluation: `max_depth = 3, min_samples_leaf = .05`.

Table 7.3: Nested CV performance (stress vs. neutral) (%).

Metric	F1 Score (SD)	Accuracy	Precision	Recall	AUC	MCC
LR	41.5 (11)	53.7	41.0	52.6	75.0	13.3
DT	63.5 (8.2)	58.4	56.0	81.0	64.8	47.2
KNN	59.1 (9.4)	59.7	61.6	63.8	63.3	28.4
NB	29.7 (11)	49.6	28.9	39.2	76.3	3.60
RFC	67.3 (6.1)	60.8	60.3	84.1	65.2	56.7
SVM	56.4 (10)	56.9	53.0	65.9	60.9	23.5

Generalisability Assessment

Two publicly available datasets were combined to act as an independent dataset for the assessment of ML algorithm generalisability. As the duration of the

300-s ST segments obtained from the independent datasets was too lengthy to support the incorporation of the ML approach into real-time systems, the signals were further divided into 60-s UST segments. Accordingly, the testing process was independently performed on both ST and UST segments.

The confusion matrix results of the ST and UST datasets are presented in Figure 7.5, and the performance metrics are summarised in Table 7.4. The F1 scores of the ST and UST were 70.7% and 61.2%, respectively. In addition, the MCC of the ST (57.9%) was higher than UST (34.4%). Moreover, the results showed that the RFC in the ST dataset correctly classified all neutral instances (100% precision), but it misclassified approximately half of the stress instances (54.7% recall). However, the performance decreased in the UST dataset, with a precision of 81.3% and a recall of 49.1%.

Figure 7.5: Confusion Matrix of the Independent Dataset (Stress vs. Neutral)

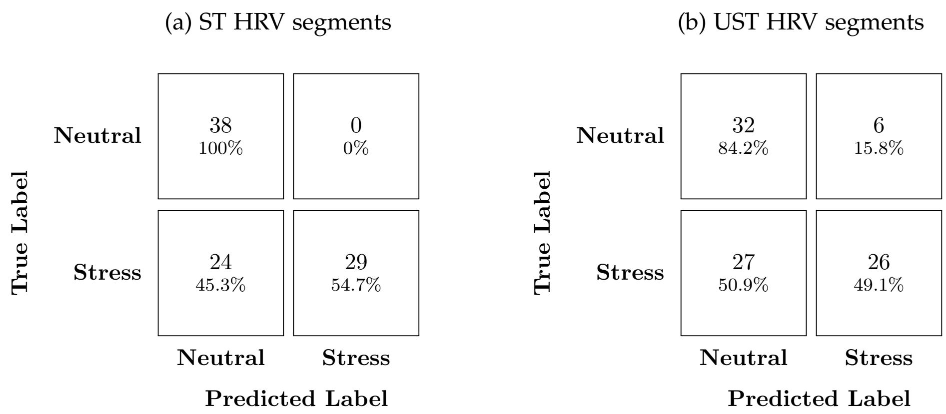


Table 7.4: Performance Evaluation of RFC on the Independent Dataset (Stress vs. Neutral) (%)

	F1 Score	Accuracy	Precision	Recall	AUC	MCC
ST (300 s)	70.7	73.6	100	54.7	70.9	57.9
UST (60 s)	61.2	63.7	81.3	49.1	63.1	34.4

7.5.2 Classification of Stress and Relaxed States

Model Selection

Table 7.5 shows the average performance metrics of the supervised ML algorithms resulting from the nested CV approach used to classify stress and relaxed instances from the original dataset.

Overall, the ML models had relatively high classification accuracy rates, ranging from 78% to 82%. RFC and DT had the highest recall score ranging from 90%-93%, indicating good performance in identifying stress instances (i.e., low false negatives). Moreover, the precision of all classifiers was in the range of 76%-86%, indicating lower false positives in stress vs. relaxed compared to stress vs. neutral. The AUC score and MCC of all classifiers were greater than 76%. Among all the classifiers, RFC had the best F1 score and stability at 82.2% and 6.3%, respectively. Thus, the RFC was selected as the best model with the following hyperparameters for the independent dataset evaluation: `max_depth = 2`, `min_samples_leaf = .05`.

Table 7.5: Nested CV Performance (Stress vs. Relax) (%)

Metric	F1 Score (SD)	Accuracy	Precision	Recall	AUC	MCC
LR	78.7 (8.6)	80.8	86.0	79.7	76.9	78.6
DT	81.5 (7.2)	76.3	76.8	92.7	82.4	81.5
KNN	78.8 (8.1)	78.2	80.9	82.3	78.7	78.9
NB	78.6 (8.9)	79.7	85.1	80.6	80.0	78.7
RFC	82.2 (6.3)	80.0	80.1	90.5	84.9	82.2
SVM	79.0 (8.6)	78.0	80.2	84.1	80.6	78.9

Generalisability Assessment

A similar procedure to the stress vs. neutral classification was followed for the generalisability assessment. The confusion matrices for the ST and UST datasets are shown in Figure 7.6, and the ML performance metrics are summarised in Table 7.6. The overall performance levels of both datasets were similar, showing an 86.6% accuracy rate for the ST dataset and 85.4% for the UST. However, the ML model performed better in the ST dataset (97.7% precision) compared to the UST dataset (91.1% precision) based on the confusion matrix and precision scores. The F1 scores for the ST and UST datasets were 88.7% and 88.2%, respectively. Moreover, the MCC of the ST (74.5%) was slightly higher than UST (69.3%).

Figure 7.6: Confusion Matrix of the Independent Dataset (Stress vs. Relaxed)

(a) ST HRV segments (b) UST HRV segments

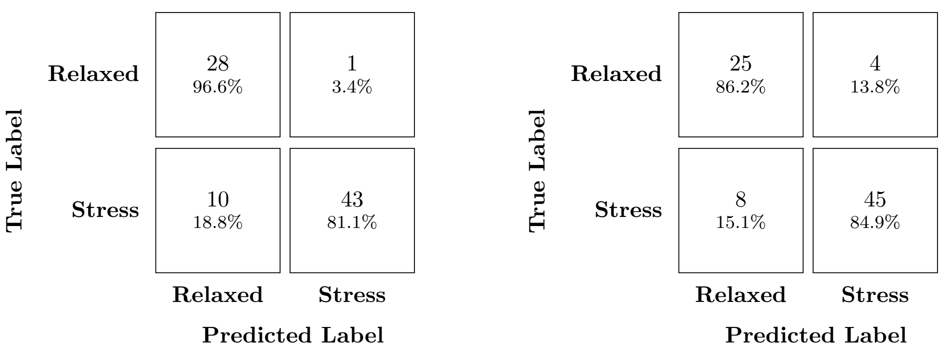


Table 7.6: Performance Evaluation of the RFC on the Independent Dataset (Stress vs. Relax) (%)

	F1 Score	Accuracy	Precision	Recall	AUC	MCC
ST (300 s)	88.7	86.6	97.7	81.1	90.4	74.5
UST (60 s)	88.2	85.4	91.8	84.9	92.1	69.3

7.6 Effects of Validation Strategy on Model Performance

To demonstrate the effect of the chosen validation strategy on classification performance, all ML models were evaluated using the standard K-fold, nested K-fold, standard LOGO, and nested LOGO CV methods, where K was set to 10. Figure 7.7 illustrates the classification performance of the stress vs. relaxed dataset using the accuracy metric. All HRV features were included in this analysis: MeanRR, RMSSD, SDNN, pNN50, LF power, HF power, LF/HF, and total power.

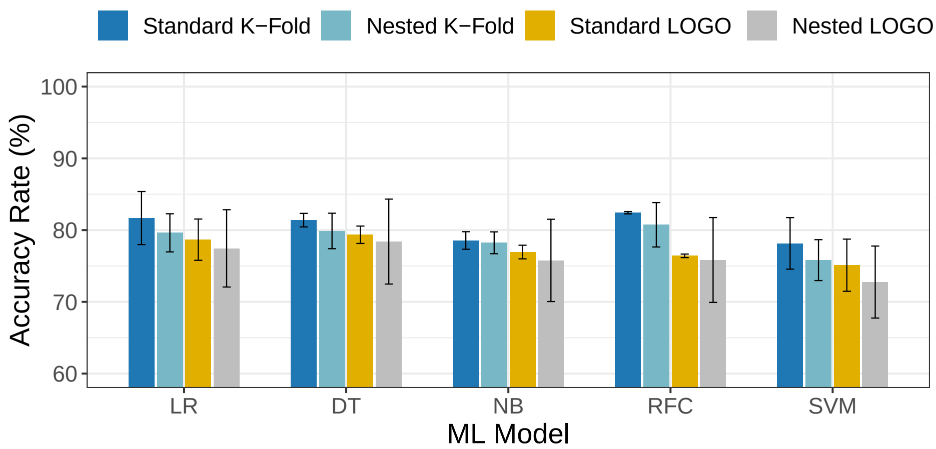


Figure 7.7: Average Accuracy Rate for each CV Method

Overall, both K-fold CV methods had higher performance than the LOGO CV methods across all ML models. The nested LOGO CV had the lowest performance among all investigated CV methods. On average, the standard K-fold performance was around 5% greater than the nested LOGO CV performance across all ML models. In particular, the RFC using the standard K-fold method was approximately 6.6% higher than that obtained using the nested LOGO CV. However, the estimated SD of the nested LOGO CV was higher than that of the remaining CV methods, which could suggest model instability.

ML performance was further evaluated in 30 trials of the LR classifier using the standard and nested versions of the K-fold and LOGO CV methods (see Figure 7.8). Each trial was computed by shuffling the observations and changing the seed parameter of the K-fold method. However, group randomisation or shuffling was not necessary as all observations in the LOGO CV were involved in the analysis, regardless of order; thus, performance was the same across all trials, as represented by a flat line in Figure 7.8.

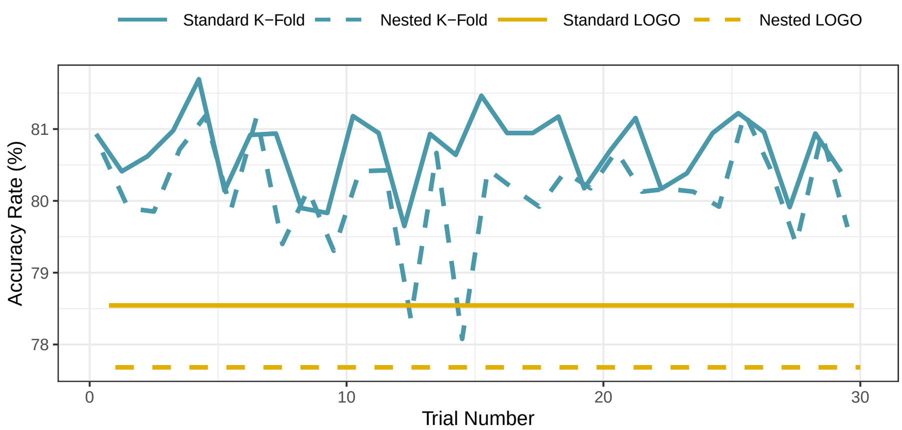


Figure 7.8: Performance of standard and nested implementations of K-fold and LOGO CV methods over 30 trials.
Note. Code Adapted from Sci-kit Learn (Scikit-Learn Developers, 2019)

GridSearchCV was used to execute hyperparameter selection for the nested methods in the inner loops and standard methods in the main loops. Next, the optimal hyperparameters were selected to fit the training set. For the nested approach, the accuracy rate was averaged on the hold-out dataset splits in the outer loop. Generally, the standard (i.e., non-nested) approaches of both CV methods had higher accuracy rates than their corresponding nested implementations. Moreover, both K-fold methods outperformed the LOGO methods.

These performance estimates are indicative of bias level in the standard CV approaches, with overestimation potentially representing higher bias due to 1) the use of the same test set for hyperparameter selection and model evaluation in the case of standard CV methods or 2) the presence of dependent observations in the training and testing sets in the case of the K-fold CV methods.

7.7 Discussion

The purpose of this study was to assess the classification performance of supervised learning algorithms in the detection of stress levels based on HRV features. This was accomplished by developing and implementing methodologically robust ML classifiers that address the key limitations associated with overfitting, overly optimistic performance, and generalisability issues identified in the current literature

Two independent binary classifiers were implemented to identify stress from non-stress states (i.e., neutral and relaxed states). Based on the nested CV model selection results, the RFC achieved the highest performance among the remaining ML algorithms in terms of identifying both stress and non-stress states. In a seminal investigation of the performance of various ML classifiers, Fernández-Delgado et al. (2014) assessed 179 classifiers from 17 families in 121 datasets and concluded that RFC had the best performance. When deploying affect recognition systems in real-world applications, explainable and interpretable ML models are important considerations for use by clinicians or individuals (Adadi & Berrada, 2018; Du et al., 2020). Given that RFC is based on ensemble learning of numerous decision trees, there may be a lack of understanding regarding how particular decisions were made between the predictors and the outcome (Aria et al., 2021). Therefore, several strategies were proposed to address this issue, including the introduction of a taxonomy of

RFC interpretative models via model visualisation and post-hoc explanatory methods (Aria et al., 2021; Haddouchi & Berrado, 2019). According to the findings of the current study, DT achieved comparable performance to RFC (see Tables 7.3 and 7.5), which is considered as a simple and easy-to-understand classification algorithm in the healthcare field (Podgorelec et al., 2002).

Generally, the RFC classification performance of stress vs. relaxed (F1 score = 82.2%) was better than stress vs. neutral (F1 score = 67.3%) due to the remarkable physiological effect of paced breathing on cardiovascular activity compared to the mild effect of mental stress tasks. The relevant HRV features used in the stress vs. relaxed classifier were significantly different between the two states. However, a note of caution is needed here as the relaxed state was associated with the paced breathing exercise itself. A better representation of the relaxed state could be generated by measuring HRV following paced breathing, similar to the approach taken by Dalmeida and Masala (2021).

The accuracy rate of the RFC in the stress vs. neutral case was 60.8%, which is significantly lower than the accuracy rate of 80% or greater reported by similar stress classification studies (Bobade & Vani, 2020; Can et al., 2020; Schmidt et al., 2018). This discrepancy in performance could be attributed to two reasons: 1) the implementation of the overlapping segmentation approach with an inadequate validation strategy, which violates the independence constraint among observations, and 2) inclusion of a large number of features relative to the size of the dataset. These issues were mitigated in a stress classification study conducted by Castaldo et al. (2019), where no segmentation was implemented for the training set and a minimal number of relevant features was selected for model development. Their findings revealed that the best classifier was KNN, with an accuracy rate of 94% and an AUC of 99%. These rates are far higher than those obtained by the current study using the RFC (60.8% accuracy, 65.2%

AUC). However, Castaldo et al. (2019) utilised a small dataset for training (25 participants) and testing (17 participants) compared to the dataset size used in the present study (training: 58 participants, testing: 38 participants). Generally, small training and testing sets do not represent the general population and, by extension, cannot support an accurate assessment of the generalisability of ML model performance (Foster et al., 2014).

To overcome performance overestimation during model selection, the nested LOGO CV method was used for the training process and hyperparameter selection, as discussed in Section 7.4.4. Despite the variance-bias trade-off (Hawkins et al., 2003), this approach is only advised for small datasets as the variance of generalisation performance can be quite high otherwise. In the case of large datasets, more than one group can be employed for validation by aggregating the participant-dependent observations to simulate the K-Fold method (e.g., leave-five-group-out).

An important aspect of ML development is generalisability. Therefore, the current study employed two independent datasets for the testing phase. Although a generalisability test evaluates how well the ML algorithms adapt to unseen data, acceptable levels of generalisation should still be determined (Futoma et al., 2020). Hence, the independent datasets were carefully selected based on the experimental protocol and HRV recording length. Nevertheless, the HRV data from these datasets were collected with an ECG-based instrument rather than a PPG-based one. Further, although the participants in the SWELL dataset underwent a work-related stress task that differed slightly from the stressors used in the original dataset, the task still evoked a mental stress workload. Thus, the determination of generalisability was not only focused on testing the ML models on the external dataset but also on extending their application to different instruments and mental stressors. Altogether, the classification performance of the RFC model on the independent datasets was relatively high, with F1 scores of 70% and 87% for the stress vs. neutral and stress vs. relaxed states, respectively.

Stress prediction based on UST HRV segments facilitates deployment in wearable devices for the purposes of improving monitoring and diagnosis. The ST classification performance metrics in the stress vs. neutral case were higher than those obtained with the UST approach. These findings are consistent with the literature (Castaldo et al., 2019; Tervonen et al., 2021), indicating that supervised learning algorithms perform better with longer HRV segments. However, no differences in performance metrics were noticeable between ST and UST HRV in the stress vs. relaxed case. These results may be due to the stronger correlation of UST HRV segments during paced breathing (i.e., relaxed state) compared to stress or neutral states, as thoroughly discussed in Chapter 5 (see Section 5.6).

Overall, performance overestimation was conveyed via the comparison of different validation strategies. Consistent with the literature, this study found that the LOGO CV and, in particular, nested LOGO CV methods obtained unbiased performance estimates compared to the standard and nested K-fold CV methods, with a mean difference of 5% among the investigated ML models. Using accelerometer data to examine various validation strategies for human activity recognition systems, Bragança et al. (2022) found that the K-fold CV of an RFC overestimated accuracy by 13% compared to the leave-one-out CV, leading to an inaccurate ML performance assessment.

7.8 Limitations

Although the present study successfully demonstrated the impact of using a robust ML methodology for small datasets, it features certain limitations in terms of dependency, labelling strategy, and model stability. First, pure dependency is not necessarily implied when the violation of the independence assumption is mitigated by grouping associated segments via the LOGO CV method (Little et al., 2017). The observations were still interdependent within a group because they were generated from the same participant. Second, the observations were assigned to one of three classes (neutral, stress, and relaxed) based on the conditions under which the data were collected. In accordance with the methods employed in similar studies (Chen et al., 2021; Coutts et al., 2020; Petrescu et al., 2021), it may have been more effective to supplement the dataset with the subjective scores reported by participants as these reflected their current stress or relaxation levels. Lastly, the relatively high SD of the outer CV performance indicates stability issues in the LOGO CV methods (see Section 7.6). Hence, further research is needed to investigate the causes of model instability and explore approaches to better stabilise the model.

7.9 Chapter Summary

This chapter addresses SRQ4 by exploring the potential for recognising stress levels from HRV measures using binary classification through the implementation of supervised learning algorithms. It begins by reviewing the existing methodological shortcomings associated with limited datasets in the relevant literature, such as data segmentation, feature selection, and model evaluation. These shortcomings pose problems for the development of robust ML algorithms, including overfitting, overly optimistic performance, and generalisability issues. Next, this chapter proposes mitigation strategies, including an appropriate selection of the validation strategy and relevant features based on the context of the investigated problem.

By following these recommendations, this study was able to identify stress from non-stress states (i.e., neutral and relaxed), with the RFC achieving the best performance, F1 scores were 67% and 82% for neutral and relaxed states, respectively. Furthermore, the generalisability aspect was explored by evaluating the RFC on the independent datasets, F1 scores were 70% and 87% for neutral and relaxed states, respectively. However, the classification performance was relatively lower than those reported in similar studies, indicating performance overestimation that most likely arose from data leakage or selection bias in the reviewed studies. When it comes to ML development using limited physiological datasets, the appropriate methodological procedures should be followed not only to improve model generalisability but also to facilitate the interpretability of the developed model based on the context of the targeted application.

CHAPTER 8

Conclusions

This research draws on a quantitative paradigm to investigate the potential of promoting mental well-being in two ways: 1) examining affective states and physiological changes in response to heart rate variability biofeedback and 2) exploring a robust approach for the development of effective stress recognition systems. Hence, this chapter presents the conclusions of the overall thesis by summarising the primary findings of the studies conducted and highlighting key contributions in light of the research questions. Lastly, it describes the major limitations encountered during data collection and analysis and provides directions for future research.

8.1 Reflections on the Research Questions

The overall aim of this research is to provide means to promote mental well-being by improving HRV analysis for eventual deployment in real-time affect recognition systems. This research investigates the short-term effects of HRVB via a paced breathing exercise across a range of affective states. The following main research question was posed:

> **RQ:** How does a single HRVB session using paced breathing mediate physiological responses across a range of affective states, and can these affects be robustly recognised by supervised learning algorithms?

For a comprehensive articulation of the main research question, four areas of focus emerged: HRV preprocessing, UST analysis, HRVB, and stress recognition.

Each area was addressed by an associated research sub-question. Accordingly, the following subsections discuss the contributions and key findings of this research through reflection on the research questions for each area.

Preprocessing of Heart Rate Variability Data

Assuring high-quality HRV signals and flexible implementations for HRV analysis is critical for the development of real-time applications as well as batch processing; hence, Chapter 4 addresses the following research sub-question:

> **SRQ1:** What signal preprocessing algorithms are necessary for a reliable real-time HRV analysis?

To address this question, Chapter 4 provides a comprehensive overview of HRV signal preprocessing algorithms and their importance in reliable HRV data analysis. Using common preprocessing methods, the existing artefacts in the HRV signal were identified via fixed and adaptive threshold detection techniques and cleaned via correction algorithms, including deletion, averaging, and interpolation. Following the correction process, the HRV measures obtained from the cleaned signals were compared to the unfiltered signal. The results demonstrate the necessity of HRV signal filtering, as discussed in Section 4.6.

More importantly, Chapter 4 describes the development of a flexible signal preprocessing algorithm based on a controllable window approach to facilitate deployment in real-time systems. The flexibility feature of the preprocessing algorithm was designed to control the necessary parameters for batch processing and real-time analysis (e.g., changing the threshold value, controlling the window size, selecting the desired preprocessing algorithm) while maintaining high-agreement levels with the Kubios HRV application. Accordingly, the outcomes of this study serve as the basis for the HRV analysis performed in the subsequent chapters. The findings of this study suggest a trade-off between artefact detection accuracy and processing time. In light of these findings, the adaptive threshold approach is recommended for offline batch processing given its high performance with respect to accuracy. However, the median-fixed threshold for detection may provide acceptable results for real-time analysis based on accuracy and processing time.

To date, the research has evaluated HRV filtering and analysis for real-time purposes by simulating both procedures using an overlapping segmentation approach with offline processing of existing datasets. Hence, this study extends the previous work by integrating the filtering component with an online HRV data acquisition framework via a BLE-based sensor using an open-source implementation (see Section 4.5). The source codes for the preprocessing algorithms and real-time HRV framework are available on GitHub (Bahameish, 2019, see Sections 4.3.3 and 4.5). To further this research, an empirical evaluation of the integrated online framework should be conducted to assess the performance and reliability of HRV filtering in a real-time setting.

Ultra-Short-Term Analysis of Heart Rate Variability Data

To simulate real-time HRV data acquisition and facilitate deployment in real-time applications, UST segments have been used to analyse HRV signals in periods of less than 5 min. Thus, this study examined the necessary requirements for UST analysis under resting and non-resting conditions, as outlined by the following research sub-question and addressed in Chapter 5:

> **SRQ2:** What are the requirements for a reliable real-time HRV analysis using UST segments under resting, stress, and paced breathing conditions?

Chapter 5 provides new insights regarding the impact of mental stress and paced breathing on HRV measures derived from short-term and UST segments and analysed using time-domain, frequency-domain, and non-linear methods (see Section 5.8). Although UST analysis has been studied in the literature, the majority of studies have focused on the assessment of HRV measures in a resting state. Moreover, a number of studies have drawn their conclusions based on an inadequate analytical test (e.g., group-mean differences) rather than the assessment of the limits of agreements between the HRV measures derived from the UST segment and 5-min interval.

Using a concurrent validity assessment of the standard 5-min HRV interval, Chapter 5 establishes the minimum reliable segment for HRV analysis based on the conditions under which the data were acquired. UST reliability was confirmed using the correlation analysis, limits of agreement, and trend consistency (see Section 5.6). Overall, the results indicate that 10 s is a reliable window for estimating MeanRR in all investigated conditions. Moreover, SDNN was reliable at 30 s in paced breathing compared to 60 s in resting and stress conditions given the periodicity of the HRV signal under paced breathing.

In addition, this study extends investigations of the influence of stress and paced breathing on HRV measures by examining the trend consistency of UST segments compared to baseline. These findings have important implications for the design and development of HRV-related real-time applications in which measurement conditions must be taken into consideration. Thus, the outcomes of this study informed the feature selection employed in the stress recognition study described in Chapter 7. To develop a full picture of the impact of mental stress on vagal tone, additional studies are needed to explore stress tasks with different stress types, higher intensity levels, and longer durations compared to the task employed in this study (i.e., TSST).

Impact of Heart Rate Variability Biofeedback on Affective States

Following the necessary preparations for HRV signal quality and an examination of the influence of stress and paced breathing on HRV measures, Chapter 6 takes a major step towards addressing the main research question by assessing the impact of a single HRVB session on a range of affective states and physiological measures, as reflected by the following research sub-question:

> **SRQ3:** What is the effect of a single paced breathing session on affective states (cognition, relaxation, stress) and physiological responses (HRV and BP)?

Chapter 6 presents an investigation of the influence of HRVB on physiological measures, psychological measures, and cognitive performance using a quantitative RCT, which allowed for the statistical assessment of group-mean differences as well as the feasibility of predicting mental well-being via HRV measures (see Section 6.6).

Further, Chapter 6 provides evidence-based knowledge regarding the impact of a single short-term HRVB session on affective states and physiological measures, both during and after the session (see Section 6.8). As demonstrated by the correct responses and attentiveness scores, cognitive performance showed promising improvement following the biofeedback intervention, despite the lack of an associated increase in vagal tone reflected by RMSSD. The lack of increase in vagal tone could be explained by several factors, such as the biofeedback protocol, duration of the biofeedback session, or participants' lack of familiarity with paced breathing.

An additional prominent finding is the improvement in relaxation levels measured via self-reported serenity scores after the biofeedback intervention. Moreover, the HRVB group demonstrated lower SBP reactivity to the stress task than the CTRL. Although HRVB had short-term effects on the HRV measures during the paced breathing intervention, only MeanRR was significant afterwards. Further, fatigue, mood, and stress were not found to be associated with HRV measures, and there were no significant differences among these affective states during the various conditions of the experiment. These outcomes could be due to the use of a single-item Likert scale for assessing mood and stress as well as the other anticipated limitations related to the study design (see Sections 6.7 and 8.3).

Despite the fact that this study employed a different stress task (N-back) than the previous study described in Chapter 5 (TSST), parasympathetically related HRV measures (i.e., RMSSD and HF power) were not significantly different during the stress task compared to baseline. In addition, these two HRV measures did not differ significantly between HRVB and CTRL groups across all time points. However, vagal tone, as indexed by RMSSD, was higher in the HRVB group post-intervention compared to baseline, with a medium effect size. To gain an in-depth understanding of vagal tone within the context of the neurovisceral integration model, future studies should employ alternative cognitive stress tasks to elicit a higher mental workload (e.g., the dual N-back task, which combines auditory and visual stimuli). With respect to the effects of paced breathing on vagal tone, the biofeedback protocol can be similarly improved by determining the RF for each participant or incorporating a longer paced breathing session.

Nonetheless, the findings of this study have significant implications regarding the impact of a single short-term HRVB session on BP, cognitive performance, HRV measures, and relaxation, thereby laying the foundation for future research (see Section 8.4).

Robust Techniques for Stress Recognition

Finally, the potential of identifying stress levels from HRV data using robust supervised learning algorithms was examined in Chapter 7 by leveraging the dataset and key findings of the previous two chapters, thereby facilitating deployment in real-time systems. The associated research sub-question was posited as follows:

> **SRQ4:** How can robust supervised learning algorithms recognise stress and relaxed states for eventual deployment in real-time systems?

Chapter 7 proposes robust strategies for limited dataset sizes in supervised learning algorithms by highlighting the current methodological limitations described in the domain-relevant reviewed studies; the limitations and strategies are explicated in Sections 7.2.1 and 7.2.2, respectively. Specifically, this proposal addresses the selection of appropriate ML methodological decisions (e.g., data segmentation, feature selection, model evaluation) when using short-term and UST HRV data in the domain of affective recognition, particularly stress recognition. Further, Chapter 7 offers an evaluation of binary stress classifiers via the proposed strategies to mitigate the shortcomings associated with limited datasets while assessing for model generalisability, as explained in Section 7.7.

Using short-term HRV data, the RFC, which had the best performance, achieved good performance in identifying stress from neutral and relaxed states (F1 scores > 70%). For deployment in real-time applications, the RFC was also tested on UST HRV segments. The results were comparable to those of the short-term segments, indicating the potential efficacy of this classifier in the development of real-time recognition systems. However, the results of the ML model development were relatively low compared to the reported performance metrics in the reviewed studies, which is indicative of overly optimistic performance estimates resulting from inappropriate validation strategies or the use of a limited dataset with dependent observations in the reviewed studies.

Another possible explanation for the notable variation in ML algorithm performance in the affective computing field is that current implementations do not consider differences between individuals in terms of perception or reaction. The individual's susceptibility to stress varies depending on a variety of factors, including genetic predisposition, personality traits, and social support (Dumitru & Cozman, 2012; Salleh, 2008). Given that a single classifier cannot adapt to all individuals, a personalisation approach based on individual differences and transfer learning has been proposed to provide a unique experience (see Section 8.4 for further details; Taylor et al., 2015; Umematsu et al., 2020; Zheng & Lu, 2016).

Overall Research Reflection

It is evident how the main research question was naturally divided into four sub-questions, and the previous subsections demonstrate how the research addressed each of these questions. Overall, it can be concluded that paced breathing during a single HRVB session does mediate physiological responses and that this mediation can be measured across a range of affective states. Furthermore, it has been shown that these affects can be robustly recognised by supervised learning algorithms. However, careful attention must be given to the methodology used when deploying such algorithms to obtain reliable and interpretable results.

8.2 Summary of Contributions

This research work contributes towards the improvement of mental well-being by using HRV data to examine changes in affective states and physiological responses as well as exploring robust ML techniques to identify stress levels. This section presents a summary of the contributions of this research to the current state of knowledge of HRV analysis within the context of affect recognition.

- An improved understanding of high-quality filtering techniques for HRV data that meet the requirements of real-time applications via the integration of a flexible open-source implementation, automatic preprocessing algorithms, and a real-time HRV data acquisition framework using BLE-based sensors.

- New insights regarding the minimum reliable window for HRV analysis based on the conditions under which the data were acquired (i.e., resting, stress, and paced breathing). The findings are underpinned by a concurrent validity assessment to facilitate future deployment in real-time systems.

- Evidence-based knowledge regarding the influence of the HRVB paced breathing intervention on affective states, executive function, and physiological responses (i.e., HRV and BP), with the findings obtained using a quantitative RCT to facilitate the statistical assessment of group-mean differences as well as the feasibility of predicting affective states via HRV measures.

- A proposal for robust ML affect recognition strategies using physiological measures that address the current limitations associated with limited datasets in the research. With an emphasis on stress as a prominent affective state, these strategies were employed in the development and evaluation of binary classifiers to predict stress levels from short-term and UST HRV data.

8.3 Limitations

The primary limitations of this research are related to population and sampling: specifically, population characteristics, sample selection, and sample size. In addition, the technological instrument used for HRV data collection may have yielded reduced accuracy due to potential bias. Specific limitations for each study were explained in the limitations section of the relevant chapters (see Sections 4.7, 5.7, 6.7 and 7.8).

Population and Sample

As discussed in Chapter 5, participants in the HRV exploratory study were students ranging from 20-36 years of age; further, all participants were recruited via a call for participation made to the university computer science department. Thus, there is a limit to how broadly the results can be applied to older or younger age groups given that demographic factors may have an impact on HRV. The subsequent study sought a more representative sample of the targeted population by focusing on a wider age range (23-62 years; see Chapter 6). However, both studies relied on volunteer and opportunity sampling techniques (Hayes, 2021), and participants were primarily recruited via university calls for participation and personal invitations. Despite the convenience and practicality of these sampling techniques in research conducted under time and resource constraints, the samples obtained may not be truly representative of the population as the use of volunteer participants alone can introduce unintended forms of bias (Hayes, 2021). A better strategy would be to limit the characteristics of the target population and purpose behind employing a random sampling technique (e.g., focusing on a specific age group or health condition).

The sample size was determined based on a priori power analysis with 80% statistical power, which is a widely accepted minimum power level (Cohen, 1988). Nonetheless, higher statistical power (e.g., 95%) is more desirable for drawing accurate conclusions about actual effects as it reduces the risk of Type II errors (Lakens, 2013). Moreover, it is well established that the performance of ML algorithms heavily depends on the size and quality of the dataset. Therefore, the sample size of the studies discussed in Chapters 5 and 6 in turn limited the dataset used in the development of the ML classification approach described in Chapter 7.

Heart Rate Variability Sensor

Another possible limitation is the use of a PPG-based sensor for HRV data collection. Previous research has revealed that motion artefacts, skin tones, and environmental noise are the main sources of PPG data inaccuracies (Castaneda et al., 2018). Given that optical sensors rely on the transmission of green infrared light, the accuracy of PPG data may decrease for highly pigmented skin tones as melanin absorbs a significant amount of green light. In a comprehensive assessment of PPG-based wrist sensors under different conditions, Bent et al. (2020) found no significant differences in HR and HRV accuracy among different skin tones. However, the accuracy and consistency of HR measurements were dependent on the type of activity and model of the sensor device. In addition, the authors reported higher error rates in PPG compared to ECG during a typing task. Interestingly, the error rates of PPG during deep breathing were lower than in all other conditions.

Therefore, further research is required to systematically validate the relevance of these findings on finger-worn PPG sensors, particularly CorSense. In this thesis, the data of a few participants were excluded due to poor signal quality, especially during the stress task. This could be attributed to unintentional hand movements, sensors not being in direct contact with the skin, or environmental noises (e.g., typing on a keyboard, vibrations from technological devices).

8.4 Future Research Directions

This research creates potential avenues for future directions to better assess the impact of HRVB on improving mental well-being.

Focused Groups

Further studies could examine the short-term effects of HRVB using paced breathing exercises on physiological measures (e.g., BP, HRV), affective states (e.g., mood, stress), or cognitive performance in real life rather than a laboratory setting, such as academic performance assessment for university students or work-related stress environment assessment for employees. Additional investigations could explore the impact of vagal tone improvement via long-term, multiple-session HRVB on emotional adaptability, resilience, and self-regulation. Moreover, future research could focus more specifically on the short-term or long-term effects of HRVB on individuals with certain diagnosed physical conditions (e.g., hypertension), psychiatric or mental health conditions (e.g., anxiety, depression), or neurodevelopmental conditions (e.g., autism, attention deficit hyperactivity disorder [ADHD]).

Multimodal and Multisensory Channels

One promising potential approach for enriching and diversifying physiological data representation is to explore the effectiveness of alternative sensory channels for conveying biofeedback information. Such advancements could be achieved by leveraging the potential of auditory and tactile interfaces to interpret HRV data. By placing particular emphasis on HCI design principles, visual cues could be supported or replaced, thereby enhancing feedback perception (see Section 2.3.3). In addition, future work could explore the incorporation of other medical signals (e.g., BP, respiration, skin conductance) in conjunction with HRV to promote health and well-being.

Combining instruments that gather various physiological data via different sensory modalities could help in the development of biofeedback interfaces based on an inclusive design approach. The incorporation of alternative sensory channels could alleviate the limitations of HRV data inaccuracy caused by unintentional hand movements in people with autism or ADHD as they tend to have increased psychomotor activity manifested by fidgeting or hand flapping (Groden et al., 2005). Additionally, it could assuage the limitations of PPG sensors based on green infrared light when used on people with highly pigmented skin (see Section 8.3; Bent et al., 2020).

Affective Forecasting

Another potentially fruitful avenue for future research is to leverage ML algorithms in extending the recognition approach for *forecasting* affective states using deep learning, as discussed in Chapter 7. Forecasting is a form of predicting future events based on collected historic data. Within the context of affective computing, Suhara et al. (2017) used deep learning algorithms to predict severe depressed mood based on self-reported data collected during the previous two weeks. In the same vein, Jaques et al. (2017) employed self-reported data, skin conductance, and weather information to predict stress and mood levels for the following day based on data from the present day. Accordingly, the prediction of future affective states could provide means for personal adjustments, early intervention, and risk prevention for serious mental health issues.

8.5 Concluding Remarks

Prior to the present research, the short-term effects of HRVB on mental well-being had not been thoroughly investigated in the context of real-time affect recognition system development. This research offers in-depth analyses of the methods used to obtain high-quality HRV signals, including the examination of various filtering methods employed in a flexible environment and the determination of the minimum reliable segment for HRV analysis under resting and non-resting conditions. One of the key strengths of this investigation is the variety of data used to assess mental well-being, including computer-logged task performance data, physiological responses, and self-reported affective states. Further, the ML models for stress classification were developed using robust techniques to provide effective solutions for stress recognition systems. Overall, this research lays the groundwork for the employment of HRV in real-time applications to predict affective states.

Bibliography

Abdullah, S., & Choudhury, T. (2018). Sensing Technologies for Monitoring Serious Mental Illnesses. *IEEE Multimedia*, *25*(1), 61–75. https://doi.org/10.1109/MMUL.2018.011921236

Acharya, U. R., Joseph, K. P., Kannathal, N., Lim, C. M., & Suri, J. S. (2006). Heart rate variability: A review. *Medical and Biological Engineering and Computing*, *44*(12), 1031–1051. https://doi.org/10.1007/s11517-006-0119-0

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Agelink, M. W., Malessa, R., Baumann, B., Majewski, T., Akila, F., Zeit, T., & Ziegler, D. (2001). Standardized tests of heart rate variability: normal ranges obtained from 309 healthy humans, and effects of age, gender, and heart rate. *Clinical Autonomic Research*, *11*(2), 99–108. https://doi.org/10.1007/BF02322053

Ali, A. M., Ahmed, A., Sharaf, A., Kawakami, N., Abdeldayem, S. M., & Green, J. (2017). The Arabic Version of The Depression Anxiety Stress Scale-21: Cumulative scaling and discriminant-validation testing. *Asian Journal of Psychiatry*, *30*, 56–58. https://doi.org/10.1016/J.AJP.2017.07.018

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, *28*(3). https://doi.org/10.1088/0967-3334/28/3/R01

Altman, D. G., & Bland, J. M. (1983). Measurement in Medicine: The Analysis of Method Comparison Studies. *The Statistician*, *32*(3), 307. https://doi.org/10.2307/2987937

Alvares, G. A., Quintana, D. S., Hickie, I. B., & Guastella, A. J. (2016). Autonomic nervous system dysfunction in psychiatric disorders and the impact of psychotropic medications: A systematic review and meta-analysis. *Journal of Psychiatry and Neuroscience*, *41*(2), 89–104. https://doi.org/10.1503/jpn.140217

Antelmi, I., De Paula, R. S., Shinzato, A. R., Peres, C. A., Mansur, A. J., & Grupi, C. J. (2004). Influence of age, gender, body mass index, and functional capacity on heart rate variability in a cohort of subjects without heart disease. *American Journal of Cardiology*, *93*(3), 381–385. https://doi.org/10.1016/j.amjcard.2003.09.065

Antony, M. M., Cox, B. J., Enns, M. W., Bieling, P. J., & Swinson, R. P. (1998). Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample. *Psychological Assessment*, *10*(2), 176–181. https://doi.org/10.1037/1040-3590.10.2.176

Anusha, A. S., Jose, J., Preejith, S. P., Jayaraj, J., & Mohanasankar, S. (2018). Physiological signal based work stress detection using unobtrusive sensors. *Biomedical Physics and Engineering Express*, *4*(6). https://doi.org/10.1088/2057-1976/aadbd4

Aria, M., Cuccurullo, C., & Gnasso, A. (2021). A comparison among interpretative proposals for Random Forests. *Machine Learning with Applications*, *6*, 100094. https://doi.org/10.1016/J.MLWA.2021.100094

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation - Advances in Research and Theory*, *8*(100), 47–89. https://doi.org/10.1016/S0079-7421(08)60452-1

Baek, H. J., Cho, C. H., Cho, J., & Woo, J. M. (2015). Reliability of ultra-short-term analysis as a surrogate of standard 5-min analysis of heart rate variability. *Telemedicine and e-Health*, *21*(5), 404–414. https://doi.org/10.1089/tmj.2014.0104

Bahameish, M. (2019). *CorSense BLE Interface* (Version 1.0.0). URL: https://github.com/Mar-iam/corSense

Bahameish, M., & Stockman, T. (2019a). The analysis of heart rate variability measures in ultra-short-term window segments. *EECS Research Open Day, Queen Mary University of London, UK*.

Bahameish, M., & Stockman, T. (2019b). Facilitating the control of stress levels in real-time as manifested in measures of heart rate variability. *The 1st International Conference on Visualization and Computer-Human Interaction (VisCHI), Doha, Qatar*.

Bahameish, M., & Stockman, T. (2020). Fundamental Considerations of HRV Analysis in the Development of Real-Time Biofeedback Systems. *2020 Computing in Cardiology Conference (CinC)*, *47*. https://doi.org/10.22489/CinC.2020.078

Bălan, O., Moise, G., Moldoveanu, A., Leordeanu, M., & Moldoveanu, F. (2019). Fear level classification based on emotional dimensions and machine learning techniques. *Sensors (Switzerland)*, *19*(7), 1–18. https://doi.org/10.3390/s19071738

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, *16*(5), 412–424. https://doi.org/10.1093/bioinformatics/16.5.412

Bankenahally, R., & Krovvidi, H. (2016). Autonomic nervous system: anatomy, physiology, and relevance in anaesthesia and critical care medicine. *BJA Education*, *16*(11), 381–387. https://doi.org/10.1093/bjaed/mkw011

Barupal, D. K., & Fiehn, O. (2019). Generating the Blood Exposome Database Using a Comprehensive Text Mining and Database Fusion Approach. *Environmental Health Perspectives*, *127*(9), 097008. https://doi.org/10.1289/EHP4713

Behinaein, B., Bhatti, A., Rodenburg, D., Hungler, P., & Etemad, A. (2020). *A Transformer Architecture for Stress Detection from ECG* (Vol. 1). Association for Computing Machinery. https://doi.org/10.1145/3460421.3480427

Benarroch, E. E. (1993). The Central Autonomic Network: Functional Organization, Dysfunction, and Perspective. *Mayo Clinic Proceedings*, *68*(10), 988–1001. https://doi.org/10.1016/S0025-6196(12)62272-1

Benchekroun, M., Chevallier, B., Istrate, D., Zalc, V., & Lenne, D. (2022). Preprocessing Methods for Ambulatory HRV Analysis Based on HRV Distribution, Variability and Characteristics (DVC). *Sensors*, *22*(5), 1–15. https://doi.org/10.3390/s22051984

Benchekroun, M., Chevallier, B., Zalc, V., Istrate, D., Lenne, D., & Vera, N. (2021). Analysis of the Impact of Inter-Beat-Interval Interpolation on real-time HRV Feature Estimation for e-Health Applications. *8ème Colloque en Télésanté et dispositifs biomédicaux (JETSAN 2021)*.

Benichou, T., Pereira, B., Mermillod, M., Tauveron, I., Pfabigan, D., Maqdasy, S., & Dutheil, F. (2018). Heart rate variability in type 2 diabetes mellitus: A systematic review and meta–analysis (R. A. Malik, Ed.). *PLOS ONE*, *13*(4), e0195166. https://doi.org/10.1371/journal.pone.0195166

Bent, B., Goldstein, B. A., Kibbe, W. A., & Dunn, J. P. (2020). Investigating sources of inaccuracy in wearable optical heart rate sensors. *npj Digital Medicine*, *3*(1), 18. https://doi.org/10.1038/s41746-020-0226-6

Berntson, G. G., Quigley, K. S., Jang, J. F., & Boysen, S. T. (1990). An Approach to Artifact Identification: Application to Heart Period Data. *Psychophysiology*, *27*(5), 586–598. https://doi.org/10.1111/j.1469-8986.1990.tb01982.x

Berntson, G. G., Quigley, K. S., Norman, G. J., & Lozano, D. L. (2009). Cardiovascular psychophysiology. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (3rd ed., pp. 183–216). Cambridge University Press. https://doi.org/10.1017/9781107415782.009

Berntson, G., Thomas Bigger Jr, J., Eckberg, D., Grossman, P., Kaufmann, P., Malik, M., Nagaraja, H., Porges, S., Saul, J., Stone, P., & Van Der Molen, M. (1997). Heart rate variability: Origins, methods, and interpretive caveats. *Psychophysiology*, *34*(6), 623–648. https://doi.org/10.1111/j.1469-8986.1997.tb02140.x

Binici, Z., Mouridsen, M. R., Køber, L., & Sajadieh, A. (2011). Decreased nighttime heart rate variability is associated with increased stroke risk. *Stroke*, *42*(11), 3196–3201. https://doi.org/10.1161/STROKEAHA.110.607697

Birkett, M. A. (2011). The Trier Social Stress Test Protocol for Inducing Psychological Stress. *Journal of Visualized Experiments*, 1–6. https://doi.org/10.3791/3238

Blair, C. (2017). Educating executive function. *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*(1-2), e1403. https://doi.org/10.1002/wcs.1403

Bluetooth Architectural Review Board. (2021). GATT Specification Supplement.

Blum, J., Rockstroh, C., & Göritz, A. S. (2020). Development and Pilot Test of a Virtual Reality Respiratory Biofeedback Approach. *Applied Psychophysiology Biofeedback*, *45*(3), 153–163. https://doi.org/10.1007/s10484-020-09468-x

Bobade, P., & Vani, M. (2020). Stress Detection with Machine Learning and Deep Learning using Multimodal Physiological Data. *Proceedings of the 2nd*

International Conference on Inventive Research in Computing Applications, ICIRCA 2020, (July), 51–57. https://doi.org/10.1109/ICIRCA48905.2020.9183244

Bonate, P. L. (2000). *Analysis of Pretest-Posttest Designs*. Chapman; Hall/CRC. https://doi.org/10.1201/9781420035926

Boucsein, W. (2012). *Electrodermal activity* (Vol. 9781461411260). Springer US. https://doi.org/10.1007/978-1-4614-1126-0

Bragança, H., Colonna, J. G., Oliveira, H. A. B. F., & Souto, E. (2022). How Validation Methodology Influences Human Activity Recognition Mobile Systems. *Sensors*, *22*(6), 2360. https://doi.org/10.3390/s22062360

Brown, B. B. (1977). Stress and the Art of Biofeedback. *Harper & Row*.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. *arXiv preprint*, 1–15.

Burma, J. S., Graver, S., Miutz, L. N., Macaulay, A., Copeland, P. V., & Smirl, J. D. (2021). The validity and reliability of ultra-short-term heart rate variability parameters and the influence of physiological covariates. *Journal of Applied Physiology*, *130*(6), 1848–1867. https://doi.org/10.1152/japplphysiol.00955.2020

Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research. *Psychiatry Research*. https://doi.org/10.1016/0165-1781(89)90047-4

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2009). Psychophysiological Science: Interdisciplinary Approaches to Classic Questions About the Mind. *Handbook of psychophysiology* (pp. 1–16). https://doi.org/10.1017/cbo9780511546396.001

Cairns, P. (2019). *Doing better statistics in human-computer interaction*. Cambridge University Press. https://doi.org/10.1017/9781108685139

Caldwell, Y. T., & Steffen, P. R. (2018). Adding HRV biofeedback to psychotherapy increases heart rate variability and improves the treatment of major depressive disorder. *International Journal of Psychophysiology*, *131*(December 2017), 96–101. https://doi.org/10.1016/j.ijpsycho.2018.01.001

Campbell, J., & Ehlert, U. (2012). Acute psychosocial stress: Does the emotional stress response correspond with physiological responses? *Psychoneuroendocrinology*, *37*(8), 1111–1134. https://doi.org/10.1016/j.psyneuen.2011.12.010

Can, Y. S., Chalabianloo, N., Ekiz, D., Fernandez-Alvarez, J., Riva, G., & Ersoy, C. (2020). Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory Stress Detection with Smartwatches. *IEEE Access*, *8*, 38146–38163. https://doi.org/10.1109/ACCESS.2020.2975351

Cannon, W. B. (1932). *The wisdom of the body*. W W Norton & Co.

Castaldo, R., Melillo, P., Bracale, U., Caserta, M., Triassi, M., & Pecchia, L. (2015). Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis. *Biomedical Signal Processing and Control*, *18*, 370–377. https://doi.org/10.1016/j.bspc.2015.02.012

Castaldo, R., Montesinos, L., Melillo, P., James, C., & Pecchia, L. (2019). Ultra-short term HRV features as surrogates of short term HRV: A case study on mental stress detection in real life. *BMC Medical Informatics and Decision Making*, *19*(1), 1–13. https://doi.org/10.1186/s12911-019-0742-y

Castaldo, R., Xu, W., Melillo, P., Pecchia, L., Santamaria, L., & James, C. (2016). Detection of mental stress due to oral academic examination via ultra-short-term HRV analysis. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, *2016-Octob*, 3805–3808. https://doi.org/10.1109/EMBC.2016.7591557

Castaneda, D., Esparza, A., Ghamari, M., Soltanpur, C., & Nazeran, H. (2018). A review on wearable photoplethysmography sensors and their potential future applications in health care. *International journal of biosensors & bioelectronics*, *4*(4), 195. https://doi.org/10.15406/IJBSBE.2018.04.00125

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107.

Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry*, *9*(1). https://doi.org/10.1038/s41398-019-0607-2

Chakraborty, S., Aich, S., Joo, M. I., Sain, M., & Kim, H. C. (2019). A Multichannel Convolutional Neural Network Architecture for the Detection of the State of Mind Using Physiological Signals from Wearable Devices. *Journal of Healthcare Engineering*, *2019*. https://doi.org/10.1155/2019/5397814

Chalmers, J. A., Quintana, D. S., Abbott, M. J., & Kemp, A. H. (2014). Anxiety disorders are associated with reduced heart rate variability: A meta-analysis. *Frontiers in Psychiatry*, *5*(JUL), 1–11. https://doi.org/10.3389/fpsyt.2014.00080

Chan, R. C., Shum, D., Toulopoulou, T., & Chen, E. Y. (2008). Assessment of executive functions: Review of instruments and identification of critical issues. *Archives of Clinical Neuropsychology*, *23*(2), 201–216. https://doi.org/10.1016/j.acn.2007.08.010

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chen, W., Zheng, S., & Sun, X. (2021). *Introducing MDPSD, a Multimodal Dataset for Psychological Stress Detection* (Vol. 1320 CCIS). Springer Singapore. https://doi.org/10.1007/978-981-16-0705-9{_}5

Cho, D., Ham, J., Oh, J., Park, J., Kim, S., Lee, N. K., & Lee, B. (2017). Detection of stress levels from biosignals measured in virtual reality environments using a kernel-based extreme learning machine. *Sensors (Switzerland)*, *17*(10). https://doi.org/10.3390/s17102435

Choi, A., & Shin, H. (2018). Quantitative analysis of the effect of an ectopic beat on the heart rate variability in the resting condition. *Frontiers in Physiology*, *9*(JUL), 922. https://doi.org/10.3389/fphys.2018.00922

Choi, K. Y., & Ishii, H. (2020). AmbienBeat: Wrist-worn mobile tactile biofeedback for heart rate rhythmic regulation. *TEI 2020 - Proceedings of the 14th International Conference on Tangible, Embedded, and Embodied Interaction*, 17–30. https://doi.org/10.1145/3374920.3374938

Chou, E. F., Khine, M., Lockhart, T., & Soangra, R. (2021). Effects of ecg data length on heart rate variability among young healthy adults. *Sensors*, *21*(18). https://doi.org/10.3390/s21186286

Citi, L., Brown, E. N., & Barbieri, R. (2012). A Real-Time Automated Point Process Method for Detection and Correction of Erroneous and Ectopic Heartbeats. *IEEE transactions on bio-medical engineering*, *59*(10), 2828. https://doi.org/10.1109/TBME.2012.2211356

Clamor, A., Koenig, J., Thayer, J. F., & Lincoln, T. M. (2016). A randomized-controlled trial of heart rate variability biofeedback for psychotic symptoms. *Behaviour Research and Therapy*, *87*, 207–215. https://doi.org/10.1016/j.brat.2016.10.003

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge. https://doi.org/10.4324/9780203771587

Coutts, L. V., Plans, D., Brown, A. W., & Collomosse, J. (2020). Deep learning with wearable based heart rate variability for prediction of mental and general health. *Journal of Biomedical Informatics*, *112*(February), 103610. https://doi.org/10.1016/j.jbi.2020.103610

Cowley, B., Filetti, M., Lukander, K., Torniainen, J., Henelius, A., Ahonen, L., Barral, O., Kosunen, I., Valtonen, T., Huotilainen, M., Ravaja, N., & Jacucci, G. (2016). The psychophysiology primer: A guide to methods and a broad review with a focus on human-computer interaction. *Foundations and Trends in Human-Computer Interaction*, *9*(3-4), 151–308. https://doi.org/10.1561/1100000065

Craig, C. L., Marshall, A. L., Sjöström, M., Bauman, A. E., Booth, M. L., Ainsworth, B. E., Pratt, M., Ekelund, U., Yngve, A., Sallis, J. F., & Oja, P. (2003). International physical activity questionnaire: 12-Country reliability and validity. *Medicine and Science in Sports and Exercise*, *35*(8), 1381–1395. https://doi.org/10.1249/01.MSS.0000078924.61453.FB

Crosswell, A. D., & Lockwood, K. G. (2020). Best practices for stress measurement: How to measure psychological stress in health research. *Health Psychology Open*, *7*(2). https://doi.org/10.1177/2055102920933072

Curtiss, J. E., Mischoulon, D., Fisher, L. B., Cusin, C., Fedor, S., Picard, R. W., & Pedrelli, P. (2021). Rising early warning signals in affect associated with future changes in depression: A dynamical systems approach. *Psychological Medicine*. https://doi.org/10.1017/S0033291721005183

Dalmeida, K. M., & Masala, G. L. (2021). HRV Features as Viable Physiological Markers for Stress Detection Using Wearable Devices. *Sensors*, *21*(8), 2873. https://doi.org/10.3390/s21082873

Dattani, S., Ritchie, H., & Roser, M. (2021). *Mental health*. URL: https://ourworldindata.org/mental-health

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94

Davis, R. C., Arce, M. A., Tobin, K. E., Palumbo, I. M., Chmielewski, M., Megreya, A. M., & Latzman, R. D. (2020). Testing Measurement Invariance of the Positive and Negative Affect Schedule (PANAS) in American and Arab University Students. *International Journal of Mental Health and Addiction*, *20*(2), 874–887. https://doi.org/10.1007/s11469-020-00411-z

de Bruin, E. I., van der Zwan, J. E., & Bögels, S. M. (2016). A RCT Comparing Daily Mindfulness Meditations, Biofeedback Exercises, and Daily Physical Exercise on Attention Control, Executive Functioning, Mindful Awareness, Self-Compassion, and Worrying in Stressed Young Adults. *Mindfulness*, *7*(5), 1182–1192. https://doi.org/10.1007/s12671-016-0561-5

Dehghani, A., Sarbishei, O., Glatard, T., & Shihab, E. (2019). A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors. *Sensors (Switzerland)*, *19*(22), 10–12. https://doi.org/10.3390/s19225026

Deka, D., & Deka, B. (2020). Investigation on HRV Signal Dynamics for Meditative Intervention. *Advances in Intelligent Systems and Computing*, *1154*, 993–1005. https://doi.org/10.1007/978-981-15-4032-5{_}89

Demetriou, C., Ozer, B. U., & Essau, C. A. (2015). Self-Report Questionnaires. *The Encyclopedia of Clinical Psychology*, 1–6. https://doi.org/10.1002/9781118625392.wbecp507

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology*, *64*(1), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Dimsdale, J. E. (2008). Psychological Stress and Cardiovascular Disease. *Journal of the American College of Cardiology*, *51*(13), 1237–1246. https://doi.org/10.1016/j.jacc.2007.12.024

Donkin, C., Little, D. R., & Houpt, J. W. (2014). Assessing the speed-accuracy trade-off effect on the capacity of information processing. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1183–1202. https://doi.org/10.1037/a0035947

Du, M., Liu, N., & Hu, X. (2020). Techniques for interpretable machine learning. *Communications of the ACM*, *63*(1), 68–77. https://doi.org/10.1145/3359786

Dumitru, V. M., & Cozman, D. (2012). The Relationship between Stress and Personality Factors. *International Journal of the Bioflux Society*, *4*(1), 34–39.

Dunn, B. D., Galton, H. C., Morgan, R., Evans, D., Oliver, C., Meyer, M., Cusack, R., Lawrence, A. D., & Dalgleish, T. (2010). Listening to your heart: How interoception shapes emotion experience and intuitive decision making. *Psychological Science*, *21*(12), 1835–1844. https://doi.org/10.1177/0956797610389191

Edgar, J. C., Keller, J., Heller, W., Miller, G. A., & Tassinary, L. G. (2009). Psychophysiology in Research on Psychopathology. In J. T. Cacioppo, L. G. Tassinary, & G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 663–687). Cambridge University Press. https://doi.org/10.1017/cbo9780511546396.028

Egger, M., Ley, M., & Hanke, S. (2019). Emotion Recognition from Physiological Signal Analysis: A Review. *Electronic Notes in Theoretical Computer Science*, *343*, 35–55. https://doi.org/10.1016/j.entcs.2019.04.009

Elliott, G., & Moore, J. (2018). Foundations of Heart Rate Variability Course.

Elzeiny, S., & Qaraqe, M. (2020). Stress classification using photoplethysmogram-based spatial and frequency domain images. *Sensors (Switzerland)*, *20*(18), 1–19. https://doi.org/10.3390/s20185312

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, and Computers*, *28*(1), 1–11. https://doi.org/10.3758/BF03203630

Ernst, G. (2017). Heart-Rate Variability—More than Heart Beats? *Frontiers in Public Health*, *5*(September), 1–12. https://doi.org/10.3389/fpubh.2017.00240

Esco, M. R., & Flatt, A. A. (2014). Ultra-short-term heart rate variability indexes at rest and post-exercise in athletes: Evaluating the agreement with accepted recommendations. *Journal of Sports Science and Medicine*, *13*(3), 535–541.

Esterman, M., Tamber-Rosenau, B. J., Chiu, Y.-C., & Yantis, S. (2010). Avoiding non-independence in fMRI data analysis: Leave one subject out. *NeuroImage*, *50*(2), 572–576. https://doi.org/10.1016/j.neuroimage.2009.10.092

Fatisson, J., Oswald, V., & Lalonde, F. (2016). Influence diagram of physiological and environmental factors affecting heart rate variability: An extended literature overview. *Heart International*, *11*(1), e32–e40. https://doi.org/10.5301/heartint.5000232

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15*, 3133–3181.

Fink, G. (2017). Stress: Concepts, Definition and History. *Reference module in neuroscience and biobehavioral psychology* (pp. 549–555). Elsevier. https://doi.org/10.1016/b978-0-12-809324-5.02208-2

Forte, G., Favieri, F., & Casagrande, M. (2019). Heart rate variability and cognitive function: A systematic review. *Frontiers in Neuroscience*, *13*(JUL), 1–11. https://doi.org/10.3389/fnins.2019.00710

Foster, K. R., Koprowski, R., & Skufca, J. D. (2014). Machine learning, medical diagnosis, and biomedical engineering research - commentary. *BioMedical Engineering Online*, *13*(1), 1–9. https://doi.org/10.1186/1475-925X-13-94

Frank, D. L., Khorshid, L., Kiffer, J. F., Moravec, C. S., & McKee, M. G. (2010). Biofeedback in medicine: who, when, why and how? *Mental Health in Family Medicine*, *7*(2), 85.

Franzon, M., & Hugdahl, K. (1987). Effects of Speed Vs. Accuracy in Vocal Reaction Time to Visual Half-Field Presentations of Incongruent (Stroop) Color-Words. *Cortex*, *23*(4), 615–629. https://doi.org/10.1016/S0010-9452(87)80052-7

Friedli, L. (2009). Mental health, resilience and inequalities. *World Health Organization*.

Friedman, N. P., & Robbins, T. W. (2022). The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacology*, *47*(1), 72–89. https://doi.org/10.1038/s41386-021-01132-0

Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., & Celi, L. A. (2020). The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, *2*(9), e489–e492. https://doi.org/10.1016/S2589-7500(20)30186-2

Garmin. (2022). *Health Science - Stress Tracking*. URL: https://www.garmin.com/en-US/garmin-technology/health-science/stress-tracking

Gedam, S., & Paul, S. (2021). A Review on Mental Stress Detection Using Wearable Sensors and Machine Learning Techniques. *IEEE Access*, *9*, 84045–84066. https://doi.org/10.1109/ACCESS.2021.3085502

Gerritsen, R. J., & Band, G. P. (2018). Breath of Life: The Respiratory Vagal Stimulation Model of Contemplative Activity. *Frontiers in Human Neuroscience*, *12*(October), 1–25. https://doi.org/10.3389/fnhum.2018.00397

Gevirtz, R. (2013). The Promise of Heart Rate Variability Biofeedback: Evidence-Based Applications. *Biofeedback*, *41*(3), 110–120. https://doi.org/10.5298/1081-5937-41.3.01

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*, *25*(2), 141–151. https://doi.org/10.11613/BM.2015.015

Giggins, O. M., Persson, U. M. C., & Caulfield, B. (2013). Biofeedback in rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, *10*(1), 6–19. https://doi.org/10.1186/1743-0003-10-60

Gjoreski, M., Kolenik, T., Knez, T., Luštrek, M., Gams, M., Gjoreski, H., & Pejović, V. (2020). Datasets for cognitive load inference using wearable sensors and psychological traits. *Applied Sciences (Switzerland)*, *10*(11). https://doi.org/10.3390/app10113843

Goessl, V. C., Curtiss, J. E., & Hofmann, S. G. (2017). The effect of heart rate variability biofeedback training on stress and anxiety: A meta-analysis. *Psychological Medicine*, *47*(15), 2578–2586. https://doi.org/10.1017/S0033291717001003

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*, *101*(23). https://doi.org/10.1161/01.CIR.101.23.E215

Gomes, P. M. C. (2018). Development of an Open-Source Python Toolbox for Heart Rate Variability (pyHRV). *Proc. Int'l Conf. on Electrical, Electronic and Computing Engineering (IcETRAN)*, 822–828.

Gonzalez, J., Soma, H., Sekine, M., & Yu, W. (2012). Psycho-physiological assessment of a prosthetic hand sensory feedback system based on an auditory display: A preliminary study. *Journal of NeuroEngineering and Rehabilitation*, *9*(1), 1–14. https://doi.org/10.1186/1743-0003-9-33

Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry*, *28*(5-6), 367–374. https://doi.org/10.1016/J.COMPBIOLCHEM.2004.09.006

Greco, A., Lanata, A., Citi, L., Vanello, N., Valenza, G., & Scilingo, E. P. (2016). Skin Admittance Measurement for Emotion Recognition: A Study over Frequency Sweep. *Electronics*, *5*(3), 46. https://doi.org/10.3390/electronics5030046

Greene, S., Thapliyal, H., & Caban-Holt, A. (2016). A Survey of Affective Computing for Stress Detection. *IEEE Consumer Electronics Magazine*, *5*(October), 44–56.

Groden, J., Goodwin, M. S., Baron, M. G., Groden, G., Velicer, W. F., Lipsitt, L. P., Hofmann, S. G., & Plummer, B. (2005). Assessing Cardiovascular Responses to Stressors in Individuals With Autism Spectrum Disorders. *Focus on Autism and Other Developmental Disabilities*, *20*(4), 244–252. https://doi.org/10.1177/10883576050200040601

Guven, O., Eftekhar, A., Kindt, W., & Constandinou, T. G. (2016). Computationally efficient real-time interpolation algorithm for non-uniform sampled biosignals. *Healthcare Technology Letters*, *3*(2), 105–110. https://doi.org/10.1049/htl.2015.0031

Haddouchi, M., & Berrado, A. (2019). A survey of methods and tools used for interpreting Random Forest. *2019 1st International Conference on Smart Systems and Data Science (ICSSD)*, 1–6. https://doi.org/10.1109/ICSSD47982.2019.9002770

Haensel, A., Mills, P. J., Nelesen, R. A., Ziegler, M. G., & Dimsdale, J. E. (2008). The relationship between heart rate variability and inflammatory markers in cardiovascular diseases. *Psychoneuroendocrinology*, *33*(10), 1305–1312. https://doi.org/10.1016/j.psyneuen.2008.08.007

Hallman, D. M., Olsson, E. M., Von Scheéle, B., Melin, L., & Lyskov, E. (2011). Effects of heart rate variability biofeedback in subjects with stress-related chronic neck pain: A pilot study. *Applied Psychophysiology Biofeedback*, *36*(2), 71–80. https://doi.org/10.1007/s10484-011-9147-0

Hamer, M., & Steptoe, A. (2007). Association between physical fitness, parasympathetic control, and proinflammatory responses to mental stress. *Psychosomatic Medicine*, *69*(7), 660–666. https://doi.org/10.1097/PSY.0b013e318148c4c0

Hansen, A. L., Johnsen, B. H., Sollers, J. J., Stenvik, K., & Thayer, J. F. (2004). Heart rate variability and its relation to prefrontal cognitive function: The effects of training and detraining. *European Journal of Applied Physiology*, *93*(3), 263–272. https://doi.org/10.1007/s00421-004-1208-0

Hansen, A. L., Johnsen, B. H., & Thayer, J. F. (2003). Vagal influence on working memory and attention. *International Journal of Psychophysiology*, *48*(3), 263–274. https://doi.org/10.1016/S0167-8760(03)00073-4

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hasnul, M. A., Aziz, N. A. A., Alelyani, S., Mohana, M., & Aziz, A. A. (2021). Electrocardiogram-based emotion recognition systems and their applications in healthcare—a review. *Sensors*, *21*(15). https://doi.org/10.3390/s21155015

Hasson, D., & Arnetz, B. B. (2005). Validation and Findings Comparing VAS vs. Likert Scales for Psychosocial Measurements. *International Electronic Journal of Health Education*, *8*(8), 178–192.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Model Assessment and Selection. *The elements of statistical learning* (2nd ed., pp. 219–260). Springer New York.

Hawkins, D. M. (2004). The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), 1–12. https://doi.org/10.1021/ci0342472

Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences*, *43*(2), 579–586. https://doi.org/10.1021/ci025626i

Hayes, N. (2021). *Doing Psychological Research* (2nd). Open University Press.

Hazer-Rau, D., Zhang, L., & Traue, H. C. (2020). A Workflow for Affective Computing and Stress Recognition from Biosignals. *Engineering Proceedings*, 85. https://doi.org/10.3390/ecsa-7-08227

He, J., Li, K., Liao, X., Zhang, P., & Jiang, N. (2019). Real-Time Detection of Acute Cognitive Stress Using a Convolutional Neural Network from Electrocardiographic Signal. *IEEE Access*, *7*, 42710–42717. https://doi.org/10.1109/ACCESS.2019.2907076

Healey, J. A., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, *6*(2), 156–166. https://doi.org/10.1109/TITS.2005.848368

Helou, K., El Helou, N., Mahfouz, M., Mahfouz, Y., Salameh, P., & Harmouche-Karaki, M. (2017). Validity and reliability of an adapted Arabic version of the long international physical activity questionnaire. *BMC Public Health*, *18*(1). https://doi.org/10.1186/s12889-017-4599-7

Hinde, K., White, G., & Armstrong, N. (2021). Wearable devices suitable for monitoring twenty four hour heart rate variability in military populations. *Sensors (Switzerland)*, *21*(4), 1–20. https://doi.org/10.3390/s21041061

Hjortskov, N., Rissén, D., Blangsted, A. K., Fallentin, N., Lundberg, U., & Søgaard, K. (2004). The effect of mental stress on heart rate variability and blood pressure during computer work. *European Journal of Applied Physiology*, *92*(1-2), 84–89. https://doi.org/10.1007/s00421-004-1055-z

Huikuri, H. V. (1995). Heart rate variability in coronary artery disease. *Journal of Internal Medicine*, *237*(4), 349–357. https://doi.org/10.1111/j.1365-2796.1995.tb01186.x

Ihmig, F. R., Antonio Gogeascoechea, H., Neurohr-Parakenings, F., Schäfer, S. K., Lass-Hennemann, J., & Michael, T. (2020). On-line anxiety level detection from biosignals: Machine learning based on a randomized controlled trial with spider-fearful individuals. *PLoS ONE*, *15*(6), 1–20. https://doi.org/10.1371/journal.pone.0231517

Jaques, N., Taylor, S., Sano, A., & Picard, R. (2017). Predicting Tomorrow's Mood, Health, and Stress Level using Personalized Multitask Learning and Domain Adaptation. *IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*, *66*, 17–33.

Jarrin, D. C., McGrath, J. J., Giovanniello, S., Poirier, P., & Lambert, M. (2012). Measurement fidelity of heart rate variability signal processing: The devil is in the details. *International Journal of Psychophysiology*, *86*(1), 88–97. https://doi.org/10.1016/j.ijpsycho.2012.07.004

Jiang, M., Mieronkoski, R., Rahmani, A. M., Hagelberg, N., Salantera, S., & Liljeberg, P. (2017). Ultra-short-Term analysis of heart rate variability for real-Time acute pain monitoring with wearable electronics. *Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017*, *2017-January*, 1025–1032. https://doi.org/10.1109/BIBM.2017.8217798

Jiang, Y., Li, W., Hossain, M. S., Chen, M., Alelaiwi, A., & Al-Hammadi, M. (2020). A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Information Fusion*, *53*, 209–221. https://doi.org/10.1016/j.inffus.2019.06.019

Kaltsas, G. A., & Chrousos, G. P. (2007). The Neuroendocrinology of Stress. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (3rd ed., pp. 315–330). Cambridge University Press.

Kamath, M. V., & Fallen, E. L. (1995). Correction of the Heart Rate Variability Signal for Ectopics and Missing Beats. In M. Malik & A. J. Camm (Eds.), *Heart rate variability* (1st ed., pp. 75–85). Futura Pub. Co.

Karavidas, M. K., Lehrer, P. M., Vaschillo, E., Vaschillo, B., Marin, H., Buyske, S., Malinovsky, I., Radvanski, D., & Hassett, A. (2007). Preliminary results of an open label study of heart rate variability biofeedback for the treatment of major depression. *Applied Psychophysiology Biofeedback*, *32*(1), 19–30. https://doi.org/10.1007/s10484-006-9029-z

Kasaoka, S., Nakahara, T., Kawamura, Y., Tsuruta, R., & Maekawa, T. (2010). Real-time monitoring of heart rate variability in critically ill patients. *Journal of Critical Care*, *25*(2), 313–316. https://doi.org/10.1016/j.jcrc.2009.06.047

Kazuma, N., Otsuka, K., Matsuoka, I., & Murata, M. (1997). Heart rate variability during 24 hours in asthmatic children. *Chronobiology International*, *14*(6), 597–606. https://doi.org/10.3109/07420529709001450

Kemeny, M. E. (2003). The Psychobiology of Stress. *Current Directions in Psychological Science*, *12*(4), 124–129. https://doi.org/10.1111/1467-8721.01246

Kemp, A. H., & Quintana, D. S. (2013). The relationship between mental and physical health: Insights from the study of heart rate variability. *International Journal of Psychophysiology*, *89*(3), 288–296. https://doi.org/10.1016/j.ijpsycho.2013.06.018

Kemp, A. H., Quintana, D. S., Felmingham, K. L., Matthews, S., & Jelinek, H. F. (2012). Depression, comorbid anxiety disorders, and heart rate variability in physically healthy, unmedicated patients: Implications for cardiovascular risk. *PLoS ONE*, *7*(2), 1–8. https://doi.org/10.1371/journal.pone.0030777

Kennedy, L., & Parker, S. H. (2019). Biofeedback as a stress management tool: a systematic review. *Cognition, Technology and Work*, *21*(2), 161–190. https://doi.org/10.1007/s10111-018-0487-x

Khazan, I. Z. (2013). Heart Rate Variability. *The clinical handbook of biofeedback*. John Wiley & Sons, Ltd.

Kim, H. G., Cheon, E. J., Bai, D. S., Lee, Y. H., & Koo, B. H. (2018). Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry Investigation*, *15*(3), 235–245. https://doi.org/10.30773/pi.2017.08.17

Kim, H., & Lee, K. J. (2013). Heart rate variability in Alzheimer's disease and mild cognitive impairment. *Alzheimer's & Dementia*, *9*(4), P211–P212. https://doi.org/10.1016/j.jalz.2013.05.395

Kim, J., & André, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(12), 2067–2083. https://doi.org/10.1109/TPAMI.2008.26

Kim, K. H., Bang, S. W., & Kim, S. R. (2004). Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, *42*(3), 419–427. https://doi.org/10.1007/BF02344719

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, *55*(4), 352–358. https://doi.org/10.1037/h0043688

Kirschbaum, C., Pirke, K. M., & Hellhammer, D. H. (1993). The 'Trier social stress test' - A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, *28*(1-2), 76–81. https://doi.org/10.1159/000119004

Kitney, R. I., & Rompelman, O. (1980). *The Study of Heart-Rate Variability.* (Vol. 94). Oxford University Press. https://doi.org/10.7326/0003-4819-94-2-287{_}3

Kleiger, R. E., Miller, J., Bigger Jr., J., & Moss, A. J. (1987). Decreased heart rate variability and its association with increased mortality after acute myocardial infarction. *American Journal of Cardiology*, *59*(4), 256–262. https://doi.org/10.1016/0002-9149(87)90795-8

Kleiger, R. E., Stein, P. K., & Bigger, J. T. (2005). Heart rate variability: Measurement and clinical utility. *Annals of Noninvasive Electrocardiology*, *10*(1), 88–101. https://doi.org/10.1111/j.1542-474X.2005.10101.x

Koelstra, S., Member, S. S., Mü hl, C., Soleymani, M., Lee, J.-S. S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I., Member, S. S., Mühl, C., Soleymani, M., Lee, J.-S. S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., & Patras, I. (2012). DEAP: A database for emotion analysis; Using physiological signals. *IEEE Transactions on Affective Computing*, *3*(1), 18–31. https://doi.org/10.1109/T-AFFC.2011.15

Koldijk, S., Neerincx, M. A., & Kraaij, W. (2018). Detecting Work Stress in Offices by Combining Unobtrusive Sensors. *IEEE Transactions on Affective Computing*, *9*(2), 227–239. https://doi.org/10.1109/TAFFC.2016.2610975

Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M. A., & Kraaij, W. (2014). The Swell knowledge work dataset for stress and user modeling research. *ICMI 2014 - Proceedings of the 2014 International Conference on Multimodal Interaction*, 291–298. https://doi.org/10.1145/2663204.2663257

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kop, W. J., Synowski, S. J., Newell, M. E., Schmidt, L. A., Waldstein, S. R., & Fox, N. A. (2011). Autonomic nervous system reactivity to positive and negative mood induction: The role of acute psychological responses and frontal electrocortical activity. *Biological Psychology*, *86*(3), 230–238. https://doi.org/10.1016/j.biopsycho.2010.12.003

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, *84*(3), 394–421. https://doi.org/10.1016/j.biopsycho.2010.03.010

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kromenacker, B. W., Sanova, A. A., Marcus, F. I., Allen, J. J., & Lane, R. D. (2018). Vagal Mediation of Low-Frequency Heart Rate Variability during Slow Yogic Breathing. *Psychosomatic Medicine*, *80*(6), 581–587. https://doi.org/10.1097/PSY.0000000000000603

Kuo, T. B. J., Lin, T., Yang, C. C. H., Li, C.-L., Chen, C.-F., & Chou, P. (1999). Effect of aging on gender differences in neural control of heart rate.

American Journal of Physiology-Heart and Circulatory Physiology, *277*(6), H2233–H2239.

Kurniawan, H., Maslov, A. V., & Pechenizkiy, M. (2013). Stress detection from speech and Galvanic Skin Response signals. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, 209–214.

Laborde, S., Allen, M. S., Borges, U., Iskra, M., Zammit, N., You, M., Hosang, T., Mosley, E., & Dosseville, F. (2022). Psychophysiological effects of slow-paced breathing at six cycles per minute with or without heart rate variability biofeedback. *Psychophysiology*, *59*(1), 1–14. https://doi.org/10.1111/psyp.13952

Laborde, S., Hosang, T., Mosley, E., & Dosseville, F. (2019a). Influence of a 30-day slow-paced breathing intervention compared to social media use on subjective sleep quality and cardiac vagal activity. *Journal of Clinical Medicine*, *8*(2). https://doi.org/10.3390/jcm8020193

Laborde, S., Iskra, M., Zammit, N., Borges, U., You, M., Sevoz-Couche, C., & Dosseville, F. (2021). Slow-paced breathing: Influence of inhalation/exhalation ratio and of respiratory pauses on cardiac vagal activity. *Sustainability (Switzerland)*, *13*(14), 1–14. https://doi.org/10.3390/su13147775

Laborde, S., Lentes, T., Hosang, T. J., Borges, U., Mosley, E., & Dosseville, F. (2019b). Influence of slow-paced breathing on inhibition after physical exertion. *Frontiers in Psychology*, *10*(AUG), 1–15. https://doi.org/10.3389/fpsyg.2019.01923

Laborde, S., Mosley, E., & Thayer, J. F. (2017). Heart rate variability and cardiac vagal tone in psychophysiological research - Recommendations for experiment planning, data analysis, and data reporting. *Frontiers in Psychology*, *8*(FEB), 1–18. https://doi.org/10.3389/fpsyg.2017.00213

Lagos, L., Vaschillo, E., Vaschillo, B., Lehrer, P., Bates, M., & Pandina, R. (2008). Heart Rate Variability Biofeedback as a Strategy for Dealing with Competitive Anxiety : A Case Study. *Biofeedback*, *36*(3), 109–115. https://doi.org/10.5298/1081-5937-39.1.11

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(NOV), 1–12. https://doi.org/10.3389/fpsyg.2013.00863

Lang, M. (2019). Automatic Near Real-Time Outlier Detection and Correction in Cardiac Interbeat Interval Series for Heart Rate Variability Analysis: Singular Spectrum Analysis-Based Approach. *JMIR Biomedical Engineering*, *4*(1), e10740. https://doi.org/10.2196/10740

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 39–58.

Lazar, J., Feng, J., & Hochheiser, H. (2017). *Research methods in human-computer interaction*. Morgan Kaufmann.

Lee, E. S., Shin, H. C., Lee, J. H., Yang, Y. J., Cho, J. J., Ahn, G., Yoon, Y. S., & Sung, E. (2015). Development of the Perceived Stress Inventory: A new questionnaire for Korean population surveys. *Korean Journal of Family Medicine*, *36*(6), 286–293. https://doi.org/10.4082/kjfm.2015.36.6.286

Lee, S., Hwang, H. B., Park, S., Kim, S., Ha, J. H., Jang, Y., Hwang, S., Park, H.-K., Lee, J., & Kim, I. Y. (2022). Mental Stress Assessment Using Ultra Short Term HRV Analysis Based on Non-Linear Method. *Biosensors*, *12*(7), 465. https://doi.org/10.3390/bios12070465

Lees, T., Shad-Kaneez, F., Simpson, A. M., Nassif, N. T., Lin, Y., & Lal, S. (2018). Heart Rate Variability as a Biomarker for Predicting Stroke, Post-stroke Complications and Functionality. *Biomarker Insights*, *13*, 20–22. https://doi.org/10.1177/1177271918786931

Legrand, N., & Allen, M. (2022). Systole: A python package for cardiac signal synchrony and analysis. *Journal of Open Source Software*, *7*(69), 3832. https://doi.org/10.21105/joss.03832

Lehrer, P. M. (2007). Biofeedback training to increase heart rate variability. In P. Lehrer, W. Sime, & R. L. Woolfolk (Eds.), *Principles and practice of stress management* (3rd, pp. 227–248). Guilford Press.

Lehrer, P., Kaur, K., Sharma, A., Shah, K., Huseby, R., Bhavsar, J., & Zhang, Y. (2020). Heart Rate Variability Biofeedback Improves Emotional and Physical Health and Performance: A Systematic Review and Meta Analysis. *Applied Psychophysiology Biofeedback*, *45*(3), 109–129. https://doi.org/10.1007/s10484-020-09466-z

Lehrer, P., Vaschillo, B., Zucker, T., Graves, J., Katsamanis, M., Aviles, M., & Wamboldt, F. (2013). Protocol for Heart Rate Variability Biofeedback Training. *Biofeedback*, *41*(3), 98–109. https://doi.org/10.5298/1081-5937-41.3.08

Lehrer, P., Vaschillo, E., Lu, S. E., Eckberg, D., Vaschillo, B., Scardella, A., & Habib, R. (2006). Heart rate variability biofeedback: Effects of age on heart rate variability, baroreflex gain, and asthma. *Chest*. https://doi.org/10.1378/chest.129.2.278

Lehrer, P. M., & Gevirtz, R. (2014). Heart rate variability biofeedback: How and why does it work? *Frontiers in Psychology*, *5*(JUL), 1–9. https://doi.org/10.3389/fpsyg.2014.00756

Lehrer, P. M., Vaschillo, E., & Vaschillo, B. (2000). Resonant frequency biofeedback training to increase cardiac variability: Rationale and manual for training. *Applied Psychophysiology Biofeedback*, *25*(3), 177–191. https://doi.org/10.1023/A:1009554825745

Lehrer, P. M., Vaschillo, E., Vaschillo, B., Lu, S. E., Eckberg, D. L., Edelberg, R., Shih, W. J., Lin, Y., Kuusela, T. A., Tahvanainen, K. U., & Hamer, R. M. (2003). Heart rate variability biofeedback increases baroreflex gain and peak expiratory flow. *Psychosomatic Medicine*, *65*(5), 796–805. https://doi.org/10.1097/01.PSY.0000089200.81962.19

Lehrer, P. M., Vaschillo, E., Vaschillo, B., Lu, S. E., Scardella, A., Siddique, M., & Habib, R. H. (2004). Biofeedback treatment for asthma. *Chest*. https://doi.org/10.1378/chest.126.2.352

Lemaire, J. B., Wallace, J. E., Lewin, A. M., de Grood, J., & Schaefer, J. P. (2011). The effect of a biofeedback-based stress management tool on physician stress: A randomized controlled clinical trial. *Open Medicine*, *5*(4), 154–165.

Levenson, R. W. (1992). Autonomic Nervous System Differences among Emotions. *Psychological Science*, *3*(1), 23–27. https://doi.org/10.1111/j.1467-9280.1992.tb00251.x

Li, K., Rüdiger, H., & Ziemssen, T. (2019). Spectral analysis of heart rate variability: Time window matters. *Frontiers in Neurology*, *10*(MAY), 1–12. https://doi.org/10.3389/fneur.2019.00545

Liapis, A., Katsanos, C., Sotiropoulos, D., Xenos, M., & Karousos, N. (2015). Recognizing Emotions in Human Computer Interaction: Studying Stress Using Skin Conductance. In J. Abascal, S. Barbosa, M. Fetter, T. Gross, P. Palanque, & M. Winckler (Eds.), *Human-computer interaction – interact 2015* (pp. 255–262). Springer International Publishing. https://doi.org/10.1007/978-3-319-22701-6{_}18

Lin, I. M., Tai, L. Y., & Fan, S. Y. (2014). Breathing at a rate of 5.5breaths per minute with equal inhalation-to-exhalation ratio increases heart rate variability. *International Journal of Psychophysiology*, *91*(3), 206–211. https://doi.org/10.1016/j.ijpsycho.2013.12.006

Lin, I. M., Wang, S. Y., Fan, S. Y., Peper, E., Chen, S. P., & Huang, C. Y. (2020). A Single Session of Heart Rate Variability Biofeedback Produced Greater Increases in Heart Rate Variability Than Autogenic Training. *Applied Psychophysiology Biofeedback*, *45*(4), 343–350. https://doi.org/10.1007/s10484-020-09483-y

Lipponen, J. A., & Tarvainen, M. P. (2019). A robust algorithm for heart rate variability time series artefact correction using novel beat classification. *Journal of Medical Engineering and Technology*, *43*(3), 173–181. https://doi.org/10.1080/03091902.2019.1640306

Lisetti, C., Nasoz, F., LeRouge, C., Ozyer, O., & Alvarez, K. (2003). Developing multimodal intelligent affective interfaces for tele-home health care. *International Journal of Human Computer Studies*, *59*(1-2), 245–255. https://doi.org/10.1016/S1071-5819(03)00051-X

Little, M. A., Varoquaux, G., Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., & Kording, K. P. (2017). Using and understanding cross-validation strategies. Perspectives on Saeb et al. *GigaScience*, *6*(5), 1–6. https://doi.org/10.1093/gigascience/gix020

Logier, R., De Jonckheere, J., & Dassonneville, A. (2004). An efficient algorithm for R-R intervals series filtering. *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, *26 VI*, 3937–3940. https://doi.org/10.1109/iembs.2004.1404100

Lovibond, P., & Lovibond, S. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the

Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*(3), 335–343. https://doi.org/10.1016/0005-7967(94)00075-U

Luneski, A., Konstantinidis, E., & Bamidis, P. D. (2010). Affective medicine: A review of affective computing efforts in medical informatics. *Methods of Information in Medicine*, *49*(3), 207–218. https://doi.org/10.3414/ME0617

Luque-Casado, A., Perales, J. C., Cárdenas, D., & Sanabria, D. (2016). Heart rate variability and cognitive processing: The autonomic response to task demands. *Biological Psychology*, *113*, 83–90. https://doi.org/10.1016/j.biopsycho.2015.11.013

Lutfi, M. (2012). Autonomic modulations in patients with bronchial asthma based on short-term heart rate variability. *Lung India*, *29*(3), 254–258. https://doi.org/10.4103/0970-2113.99111

Mahinrad, S., Jukema, J. W., Van Heemst, D., MacFarlane, P. W., Clark, E. N., De Craen, A. J., & Sabayan, B. (2016). 10-Second heart rate variability and cognitive function in old age. *Neurology*, *86*(12), 1120–1127. https://doi.org/10.1212/WNL.0000000000002499

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. H. (2021). NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods*, *53*(4), 1689–1696. https://doi.org/10.3758/s13428-020-01516-y

Maleki, F., Muthukrishnan, N., Ovens, K., Reinhold, C., & Forghani, R. (2020). Machine Learning Algorithm Validation: From Essentials to Advanced Applications and Implications for Regulatory Certification and Deployment. *Neuroimaging Clinics of North America*, *30*(4), 433–445. https://doi.org/10.1016/j.nic.2020.08.004

Malik, M., Bigger, J. T., Camm, A. J., Kleiger, R. E., Malliani, A., Moss, A. J., & Schwartz, P. J. (1996). Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, *17*(3), 354–381. https://doi.org/10.1093/oxfordjournals.eurheartj.a014868

Malik, M., Cripps, T., Farrell, T., & Camm, A. J. (1989). Prognostic value of heart rate variability after myocardial infarction. A comparison of different data-processing methods. *Medical & Biological Engineering & Computing*, *27*(6), 603–611. https://doi.org/10.1007/BF02441642

Mather, M., & Thayer, J. F. (2018). How heart rate variability affects emotion regulation brain networks. *Current Opinion in Behavioral Sciences*, *19*, 98–104. https://doi.org/10.1016/j.cobeha.2017.12.017

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, *405*(2), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9

McDonald, A. H. (1980). Mechanisms affecting heart-rate. In R. I. Kitney & O. Rompelman (Eds.), *The study of heart-rate variability* (pp. 3–12). Oxford University Press.

McDuff, D. J., Hernandez, J., Gontarek, S., & Picard, R. W. (2016). COGCAM: Contact-free measurement of cognitive stress during computer tasks with a digital camera. *Conference on Human Factors in Computing Systems - Proceedings*, 4000–4004. https://doi.org/10.1145/2858036.2858247

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Medical Working Group. (2011a). *Heart rate profile* (tech. rep.). Bluetooth.

Medical Working Group. (2011b). *Heart rate service* (tech. rep.). Bluetooth.

Melillo, P., Bracale, M., & Pecchia, L. (2011). Nonlinear Heart Rate Variability features for real-life stress detection. Case study: Students under stress due to university examination. *BioMedical Engineering Online*, *10*(1), 96. https://doi.org/10.1186/1475-925X-10-96

Melo, H. M., Martins, T. C., Nascimento, L. M., Hoeller, A. A., Walz, R., & Takase, E. (2018). Ultra-short heart rate variability recording reliability: The effect of controlled paced breathing. *Annals of Noninvasive Electrocardiology*, *23*(5), 1–9. https://doi.org/10.1111/anec.12565

Mertz, L. (2016). The End of Seizures and Depression ? *IEEE Pulse*, (January), 2015–2017.

Meule, A., & Kübler, A. (2017). A Pilot Study on the Effects of Slow Paced Breathing on Current Food Craving. *Applied Psychophysiology Biofeedback*, *42*(1), 59–68. https://doi.org/10.1007/s10484-017-9351-7

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill.

Mouton, C., Ronson, A., Razavi, D., Delhaye, F., Kupper, N., Paesmans, M., Moreau, M., Nogaret, J. M., Hendlisz, A., & Gidron, Y. (2012). The relationship between heart rate variability and time-course of carcinoembryonic antigen in colorectal cancer. *Autonomic Neuroscience: Basic and Clinical*, *166*(1), 96–99. https://doi.org/10.1016/J.AUTNEU.2011.10.002

Munoz, M. L., Van Roon, A., Riese, H., Thio, C., Oostenbroek, E., Westrik, I., De Geus, E. J., Gansevoort, R., Lefrandt, J., Nolte, I. M., & Snieder, H. (2015). Validity of (Ultra-)Short recordings for heart rate variability measurements. *PLoS ONE*, *10*(9), 1–15. https://doi.org/10.1371/journal.pone.0138921

Muñoz Diosdado, A., Gálvez Coyt, G., & Pérez Uribe, B. M. (2010). Oscillations in the evaluation of fractal dimension of RR intervals time series. *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10*, 4570–4573. https://doi.org/10.1109/IEMBS.2010.5625941

Nakamura, Y., Yamamoto, Y., & Muraoka, I. (1993). Autonomic control of heart rate during physical exercise and fractal dimension of heart rate variability. *Journal of Applied Physiology*, *74*(2), 875–881. https://doi.org/10.1152/jappl.1993.74.2.875

Nestoriuc, Y., & Martin, A. (2007). Efficacy of biofeedback for migraine: A meta-analysis. *Pain*, *128*(1-2), 111–127. https://doi.org/10.1016/J.PAIN.2006.09.007

Nussinovitch, U., Elishkevitz, K. P., Katz, K., Nussinovitch, M., Segev, S., Volovitz, B., & Nussinovitch, N. (2011). Reliability of ultra-short ECG indices for heart rate variability. *Annals of Noninvasive Electrocardiology*, *16*(2), 117–122. https://doi.org/10.1111/j.1542-474X.2011.00417.x

Oskooei, A., Chau, S. M., Weiss, J., Sridhar, A., Martínez, M. R., & Michel, B. (2021). DeStress: Deep Learning for Unsupervised Identification of Mental Stress in Firefighters from Heart-Rate Variability (HRV) Data. *Studies in Computational Intelligence*, *914*, 93–105. https://doi.org/10.1007/978-3-030-53352-6{_}9

Papadopoulou, A., Berry, J., Knight, T., & Picard, R. (2019). Affective Sleeve: Wearable Materials with Haptic Action for Promoting Calmness. In N. Streitz & S. Konomi (Eds.), *Distributed, ambient and pervasive interactions* (pp. 304–319). Springer International Publishing. https://doi.org/10.1007/978-3-030-21935-2{_}23

Pasadyn, S. R., Soudan, M., Gillinov, M., Houghtaling, P., Phelan, D., Gillinov, N., Bittel, B., & Desai, M. Y. (2019). Accuracy of commercially available heart rate monitors in athletes: A prospective study. *Cardiovascular Diagnosis and Therapy*, *9*(4), 379–385. https://doi.org/10.21037/cdt.2019.06.05

Pecchia, L., Castaldo, R., Montesinos, L., & Melillo, P. (2018). Are ultra-short heart rate variability features good surrogates of short-term ones? State-of-the-art review and recommendations. *Healthcare Technology Letters*, *5*(3), 94–100. https://doi.org/10.1049/htl.2017.0090

Peltola, M. A. (2012). Role of editing of R–R intervals in the analysis of heart rate variability. *Frontiers in Physiology*, *3*, 1–10. https://doi.org/10.3389/fphys.2012.00148

Peng, C. K., Mietus, J. E., Liu, Y., Khalsa, G., Douglas, P. S., Benson, H., & Goldberger, A. L. (1999). Exaggerated heart rate oscillations during two meditation techniques. *International Journal of Cardiology*, *70*(2), 101–107. https://doi.org/10.1016/S0167-5273(99)00066-2

Petrescu, L., Petrescu, C., Oprea, A., Mitruț, O., Moise, G., Moldoveanu, A., & Moldoveanu, F. (2021). Machine Learning Methods for Fear Classification Based on Physiological Features. *Sensors*, *21*(13), 4519. https://doi.org/10.3390/s21134519

Picard, R. (1997). *Affective computing*. The MIT Press.

Picard, R. W. (2009). Future affective technology for autism and emotion communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3575–3584. https://doi.org/10.1098/rstb.2009.0143

Picard, R. W. (2016). Automating the Recognition of Stress and Emotion: From Lab to Real-World Impact. *IEEE MultiMedia*, *23*(3), 3–7. https://doi.org/10.1109/MMUL.2016.38

Picard, R. W., & Healey, J. (1997). Affective wearables. *Personal and Ubiquitous Computing*, *1*(4), 231–240. https://doi.org/10.1007/BF01682026

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(10), 1175–1191. https://doi.org/10.1109/34.954607

Pichot, V., Roche, F., Celle, S., Barthélémy, J. C., & Chouchou, F. (2016). HRV analysis: A free software for analyzing cardiac autonomic activity. *Frontiers in Physiology*, 1–15. https://doi.org/10.3389/fphys.2016.00557

Plews, D. J., Laursen, P. B., Stanley, J., Kilding, A. E., & Buchheit, M. (2013). Training adaptation and heart rate variability in elite endurance athletes: Opening the door to effective monitoring. *Sports Medicine*, *43*(9), 773–781. https://doi.org/10.1007/s40279-013-0071-8

Podgorelec, V., Kokol, P., Stiglic, B., & Rozman, I. (2002). Decision Trees: An Overview and Their Use in Medicine. *Journal of Medical Systems 2002 26:5*, *26*(5), 445–463. https://doi.org/10.1023/A:1016409317640

Porges, S. W. (1995). Cardiac vagal tone: A physiological index of stress. *Neuroscience and Biobehavioral Reviews*, *19*(2), 225–233. https://doi.org/10.1016/0149-7634(94)00066-A

Prinsloo, G. E., Rauch, H. G., Karpul, D., & Derman, W. E. (2013). The effect of a single session of short duration heart rate variability biofeedback on EEG: A pilot study. *Applied Psychophysiology Biofeedback*, *38*(1), 45–56. https://doi.org/10.1007/s10484-012-9207-0

Prinsloo, G. E., Rauch, H. G., Lambert, M. I., Muench, F., Noakes, T. D., & Derman, W. E. (2011). The effect of short duration heart rate variability (HRV) biofeedback on cognitive performance during laboratory induced cognitive stress. *Applied Cognitive Psychology*, *25*(5), 792–801. https://doi.org/10.1002/acp.1750

Quintana, D. S., Westlye, L. T., Kaufmann, T., Rustan, Ø. G., Brandt, C. L., Haatveit, B., Steen, N. E., & Andreassen, O. A. (2016a). Reduced heart rate variability in schizophrenia and bipolar disorder compared to healthy controls. *Acta Psychiatrica Scandinavica*, *133*(1), 44–52. https://doi.org/10.1111/acps.12498

Quintana, D. S., Alvares, G. A., & Heathers, J. A. (2016b). Guidelines for Reporting Articles on Psychiatry and Heart rate variability (GRAPH): recommendations to advance research communication. *Translational psychiatry*, *6*(5), e803. https://doi.org/10.1038/tp.2016.73

Quintana, D. S., Guastella, A. J., Outhred, T., Hickie, I. B., & Kemp, A. H. (2012). Heart rate variability is associated with emotion recognition: Direct evidence for a relationship between the autonomic nervous system and social cognition. *International Journal of Psychophysiology*, *86*(2), 168–172. https://doi.org/10.1016/j.ijpsycho.2012.08.012

Quintana, D. S., & Heathers, J. A. (2014). Considerations in the assessment of heart rate variability in biobehavioral research. *Frontiers in Psychology*, *5*(JUL), 1–10. https://doi.org/10.3389/fpsyg.2014.00805

Ramshur, J. (2010). *Design, evaluation, and application of heart rate variability analysis software (HRVAS)* (Master's thesis). The University of Memphis.

Ratcliff, R. (1993). Methods for dealing with response time outliers. *Psychological Bulletin*, *114*(3), 510–532.

Reiss, A., Indlekofer, I., Schmidt, P., & Van Laerhoven, K. (2019). Deep PPG: Large-scale heart rate estimation with convolutional neural networks. *Sensors (Switzerland)*, *19*(14), 1–27. https://doi.org/10.3390/s19143079

Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. *Computers in Biology and Medicine*, *112*. https://doi.org/10.1016/j.compbiomed.2019.103375

Ribeiro, G. d. S., Neves, V. R., Deresz, L. F., Melo, R. D., Dal Lago, P., & Karsten, M. (2018). Can RR intervals editing and selection techniques interfere with the analysis of heart rate variability? *Brazilian Journal of Physical Therapy*, *22*(5), 383–390. https://doi.org/10.1016/j.bjpt.2018.03.008

Rockstroh, C., Blum, J., & Göritz, A. S. (2019). Virtual reality in the application of heart rate variability biofeedback. *International Journal of Human Computer Studies*, *130*(June), 209–220. https://doi.org/10.1016/j.ijhcs.2019.06.011

Rubin, J., Abreu, R., Ahern, S., Eldardiry, H., & Bobrow, D. G. (2016). Time, frequency & complexity analysis for recognizing panic states from physiologic time-series. *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, *2016-May*. https://doi.org/10.4108/eai.16-5-2016.2263292

Russell, J. A. (1979). Affective space is bipolar. *Journal of Personality and Social Psychology*, *37*(3), 345–356. https://doi.org/10.1037/0022-3514.37.3.345

Russo, M. A., Santarelli, D. M., & O'Rourke, D. (2017). The physiological effects of slow breathing in the healthy human. *Breathe*, *13*(4), 298–309. https://doi.org/10.1183/20734735.009817

Sacha, J. (2013). Why should one normalize heart rate variability with respect to average heart rate. *Frontiers in Physiology*, *4*. https://doi.org/10.3389/FPHYS.2013.00306

Sacha, J., & Pluta, W. (2008). Alterations of an average heart rate change heart rate variability due to mathematical reasons. *International Journal of Cardiology*, *128*(3), 444–447. https://doi.org/10.1016/J.IJCARD.2007.06.047

Sachdev, P. S., Blacker, D., Blazer, D. G., Ganguli, M., Jeste, D. V., Paulsen, J. S., & Petersen, R. C. (2014). Classifying neurocognitive disorders: The DSM-5 approach. *Nature Reviews Neurology*, *10*(11), 634–642. https://doi.org/10.1038/nrneurol.2014.181

Salahuddin, L., Cho, J., Jeong, M. G., & Kim, D. (2007a). Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*. https://doi.org/10.1109/IEMBS.2007.4353378

Salahuddin, L., Jeong, M. G., & Kim, D. (2007b). Ultra short term analysis of heart rate variability using normal sinus rhythm and atrial fibrillation ECG data. *HEALTHCOM 2007: Ubiquitous Health in Aging Societies - 2007 9th International Conference on e-Health Networking, Application and Services*, 240–243. https://doi.org/10.1109/HEALTH.2007.381639

Salleh, M. R. (2008). Life event, stress and illness. *Malaysian Journal of Medical Sciences*, *15*(4), 9–18.

Salo, M. A., Huikuri, H. V., & Seppanen, T. (2001). Ectopic beats in heart rate variability analysis: Effects of editing on time and frequency domain measures. *Annals of Noninvasive Electrocardiology*, *6*(1), 5–17. https://doi.org/10.1111/J.1542-474X.2001.TB00080.X

Sammito, S., & Böckelmann, I. (2016). Factors influencing heart rate variability. *International Cardiovascular Forum Journal*, *6*, 18–22. https://doi.org/10.17987/icfj.v6i0.242

Sarkar, P., & Etemad, A. (2020). Self-supervised ECG Representation Learning for Emotion Recognition. *IEEE Transactions on Affective Computing*, 1–13. https://doi.org/10.1109/TAFFC.2020.3014842

Schmidt, P., Reiss, A., Duerichen, R., & Van Laerhoven, K. (2018). Introducing WeSAD, a multimodal dataset for wearable stress and affect detection. *ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction*, 400–408. https://doi.org/10.1145/3242969.3242985

Schmidt, P., Reiss, A., Dürichen, R., & Laerhoven, K. V. (2019). Wearable-based affect recognition—a review. *Sensors (Switzerland)*, *19*(19), 1–42. https://doi.org/10.3390/s19194079

Schober, P., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia and Analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Schoenberg, P. L., & David, A. S. (2014). Biofeedback for psychiatric disorders: A systematic review. *Applied Psychophysiology Biofeedback*, *39*(2), 109–135. https://doi.org/10.1007/S10484-014-9246-9/TABLES/8

Schuman, D. L., & Killian, M. O. (2019). Pilot Study of a Single Session Heart Rate Variability Biofeedback Intervention on Veterans' Posttraumatic Stress Symptoms. *Applied Psychophysiology Biofeedback*, *44*(1), 9–20. https://doi.org/10.1007/s10484-018-9415-3

Schwartz, M. S., Collura, T. F., Kamiya, J., & Schwartz, N. M. (2016). The History and Definitions of Biofeedback and Applied Psychophysiology. In M. S. Schwartz & F. Andrasik (Eds.), *Biofeedback: A practitioner's guide* (4th ed.). The Guilford Press.

Scikit-Learn Developers. (2019). *Nested versus non-nested cross-validation*. URL: https://scikit-learn.org/stable/auto_examples/model_selection/plot_nested_cross_validation_iris.html

Scikit-Learn Developers. (2022). *Nested cross-validation | Scikit-Learn Course*. URL: https://inria.github.io/scikit-learn-mooc/python_scripts/cross_validation_nested.html

Selye, H. (1976). *Stress in Health and Disease*. Butterworths.

Selye, H. (1978). *The stress of life* (Revised). McGraw-Hill.

Shaffer, F., Shearman, S., Meehan, Z., Gravett, N., & Urban, H. (2019). The promise of ultra-short-term (UST) heart rate variability measurements: a comparison of Pearson product-moment correlation coefficient and limits of agreement (LoA) concurrent validity criteria. In D. Moss & F. Shaffer (Eds.), *Physiological recording technology and applications in biofeedback and neurofeedback* (pp. 214–220). Association for Applied Psychophysiology; Biofeedback (AAPB).

Shaffer, F., & Ginsberg, J. P. (2017). An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in Public Health*, *5*(September), 1–17. https://doi.org/10.3389/fpubh.2017.00258

Shaffer, F., & Meehan, Z. M. (2020). A Practical Guide to Resonance Frequency Assessment for Heart Rate Variability Biofeedback. *Frontiers in Neuroscience*, *14*(October). https://doi.org/10.3389/fnins.2020.570400

Shaffer, F., Meehan, Z. M., & Zerr, C. L. (2020). A Critical Review of Ultra-Short-Term Heart Rate Variability Norms Research. *Frontiers in Neuroscience*, *14*(November), 1–11. https://doi.org/10.3389/fnins.2020.594880

Shaffer, F., Shearman, S., & Meehan, Z. M. (2016). The Promise of Ultra-Short-Term (UST) Heart Rate Variability Measurements. *Biofeedback*, *44*(4), 229–233. https://doi.org/10.5298/1081-5937-44.3.09

Shaffer, F., & Venner, J. (2013). Heart Rate Variability Anatomy and Physiology. *Biofeedback*, *41*(1), 13–25. https://doi.org/10.5298/1081-5937-41.1.05

Sheridan, D. C., Domingo, K. N., Dehart, R., & Baker, S. D. (2021). Heart Rate Variability Duration: Expanding the Ability of Wearable Technology to Improve Outpatient Monitoring? *Frontiers in Psychiatry*, *12*(June), 1–6. https://doi.org/10.3389/fpsyt.2021.682553

Shiraishi, Y., Katsumata, Y., Sadahiro, T., Azuma, K., Akita, K., Isobe, S., Yashima, F., Miyamoto, K., Nishiyama, T., Tamura, Y., Kimura, T., Nishiyama, N., Aizawa, Y., Fukuda, K., & Takatsuki, S. (2018). Real-time analysis of the heart rate variability during incremental exercise for the detection of the ventilatory threshold. *Journal of the American Heart Association*, *7*(1), 1–12. https://doi.org/10.1161/JAHA.117.006612

Skoluda, N., Strahler, J., Schlotz, W., Niederberger, L., Marques, S., Fischer, S., Thoma, M. V., Spoerri, C., Ehlert, U., & Nater, U. M. (2015). Intra-individual psychological and physiological responses to acute laboratory stressors of different intensity. *Psychoneuroendocrinology*, *51*, 227–236. https://doi.org/10.1016/j.psyneuen.2014.10.002

Smets, E., Casale, P., Großekathöfer, U., Lamichhane, B., De Raedt, W., Bogaerts, K., Van Diest, I., & Van Hoof, C. (2016). Comparison of Machine Learning Techniques for Psychophysiological Stress Detection. *Pervasive computing paradigms for mental health* (pp. 13–22). https://doi.org/10.1007/978-3-319-32270-4{_}2

Smith, D. L., Manning, T. S., & Petruzzello, S. J. (2001). Effect of strenuous live-fire drills on cardiovascular and psychological responses of recruit firefighters. *Ergonomics*, *44*(3), 244–254. https://doi.org/10.1080/00140130121115

Soler, A. I. R., Silva, L. E. V., Fazan, R., & Murta, L. O. (2018). The impact of artifact correction methods of RR series on heart rate variability parameters. *Journal of Applied Physiology*, *124*(3), 646–652. https://doi.org/10.1152/japplphysiol.00927.2016

Solernó, J. I., Pérez Chada, D., Guinjoan, S. M., Pérez Lloret, S., Hedderwick, A., Vidal, M. F., Cardinali, D. P., & Vigo, D. E. (2012). Cardiac autonomic activity predicts dominance in verbal over spatial reasoning tasks: Results from a preliminary study. *Autonomic Neuroscience: Basic and Clinical*, *167*(1), 78–80. https://doi.org/10.1016/J.AUTNEU.2011.10.008

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Srirampraakash, S., Prasanna, V. D., & Murthy, O. V. (2017). Stress Detection in Working People. *Procedia Computer Science*, *115*, 359–366. https://doi.org/10.1016/j.procs.2017.09.090

Starcke, K., Wiesen, C., Trotzke, P., & Brand, M. (2016). Effects of acute laboratory stress on executive functions. *Frontiers in Psychology*, *7*(MAR), 1–8. https://doi.org/10.3389/fpsyg.2016.00461

Steffen, P. R., Austin, T., DeBarros, A., & Brown, T. (2017). The Impact of Resonance Frequency Breathing on Measures of Heart Rate Variability, Blood Pressure, and Mood. *Frontiers in Public Health*, *5*(August), 6–11. https://doi.org/10.3389/fpubh.2017.00222

Steffen, P. R., Bartlett, D., Channell, R. M., Jackman, K., Cressman, M., Bills, J., & Pescatello, M. (2021). Integrating Breathing Techniques Into Psychotherapy to Improve HRV: Which Approach Is Best? *Frontiers in Psychology*, *12*(February), 1–11. https://doi.org/10.3389/fpsyg.2021.624254

Steptoe, A., & Vogele, C. (1991). Methodology of mental stress testing in cardiovascular research. *Circulation*, *83*(4 SUPPL.).

Stevens, L. M., Mortazavi, B. J., Deo, R. C., Curtis, L., & Kao, D. P. (2020). Recommendations for reporting machine learning analyses in clinical research. *Circulation: Cardiovascular Quality and Outcomes*, (October), 782–793. https://doi.org/10.1161/CIRCOUTCOMES.120.006556

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*(4), 1096–1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. *Teaching of Psychology*, *44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Strüven, A., Holzapfel, C., Stremmel, C., & Brunner, S. (2021). Obesity, nutrition and heart rate variability. *International Journal of Molecular Sciences*, *22*(8), 1–13. https://doi.org/10.3390/ijms22084215

Suhara, Y., Xu, Y., & Pentland, A. S. (2017). DeepMood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. *26th International World Wide Web Conference, WWW 2017*, 715–724. https://doi.org/10.1145/3038912.3052676

Suleiman, K. H., Yates, B. C., Berger, A. M., Pozehl, B., & Meza, J. (2010). Translating the pittsburgh sleep quality index into arabic. *Western Journal of Nursing Research*, *32*(2), 250–268. https://doi.org/10.1177/0193945909348230

Synnott, J., Dietzel, D., & Ioannou, M. (2015). A review of the polygraph: history, methodology and current status. *Crime Psychology Review*, *1*(1), 59–83. https://doi.org/10.1080/23744006.2015.1060080

Sztajzel, J. (2004). Heart rate variability: A noninvasive electrocardiographic method to measure the autonomic nervous system. *Swiss Medical Weekly*, *134*(35-36), 514–522. https://doi.org/2004/35/smw-10321

Tan, G., Dao, T. K., Farmer, L., Sutherland, R. J., & Gevirtz, R. (2011). Heart rate variability (HRV) and posttraumatic stress disorder (PTSD): A pilot study. *Applied Psychophysiology Biofeedback*, *36*(1), 27–35. https://doi.org/10.1007/s10484-010-9141-y

Tarvainen, M., Lipponen, L., Niskanen, R.-a., & Ranta-aho, P. (2016). Kubios HRV Software: User's Guide.

Tarvainen, M. P., Niskanen, J. P., Lipponen, J. A., Ranta-aho, P. O., & Karjalainen, P. A. (2014). Kubios HRV - Heart rate variability analysis software. *Computer Methods and Programs in Biomedicine*, *113*(1), 210–220. https://doi.org/10.1016/j.cmpb.2013.07.024

Taylor, S., Jaques, N., Weixuan Chen, Fedor, S., Sano, A., & Picard, R. (2015). Automatic identification of artifacts in electrodermal activity data. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, *176*(1), 1934–1937. https://doi.org/10.1109/EMBC.2015.7318762

Tervonen, J., Pettersson, K., & Mäntyjärvi, J. (2021). Ultra-short window length and feature importance analysis for cognitive load detection from wearable sensors. *Electronics (Switzerland)*, *10*(5), 1–19. https://doi.org/10.3390/electronics10050613

Thayer, J. F., Hansen, A. L., Saus-Rose, E., & Johnsen, B. H. (2009a). Heart rate variability, prefrontal neural function, and cognitive performance: The neurovisceral integration perspective on self-regulation, adaptation, and health. *Annals of Behavioral Medicine*, *37*(2), 141–153. https://doi.org/10.1007/s12160-009-9101-z

Thayer, J. F., & Lane, R. D. (2000). A model of neurovisceral integration in emotion regulation and dysregulation. *Journal of Affective Disorders*, *61*(3), 201–216. https://doi.org/10.1016/S0165-0327(00)00338-4

Thayer, J. F., Sollers, J. J., Labiner, D. M., Weinand, M., Herring, A. M., Lane, R. D., & Ahern, G. L. (2009b). Age-related differences in prefrontal control of heart rate in humans: A pharmacological blockade study. *International Journal of Psychophysiology*, *72*(1), 81–88. https://doi.org/10.1016/J.IJPSYCHO.2008.04.007

Theeng Tamang, M. R., Sharif, M. S., Al-Bayatti, A. H., Alfakeeh, A. S., & Alsayed, A. O. (2020). A machine-learning-based approach to predict the health impacts of commuting in large cities: Case study of London. *Symmetry*, *12*(5). https://doi.org/10.3390/SYM12050866

Thuraisingham, R. A. (2006). Preprocessing RR interval time series for heart rate variability analysis and estimates of standard deviation of RR intervals. *Computer Methods and Programs in Biomedicine*, *83*(1), 78–82. https://doi.org/10.1016/j.cmpb.2006.05.002

Tinello, D., Kliegel, M., & Zuber, S. (2022). Does Heart Rate Variability Biofeedback Enhance Executive Functions Across the Lifespan? A Systematic Review. *Journal of Cognitive Enhancement*, *6*(1), 126–142. https://doi.org/10.1007/s41465-021-00218-3

Tutunji, R., Kogias, N., Kapteijns, B., Krentz, M., Krause, F., Vassena, E., & Hermans, E. (2021). Using wearable biosensors and ecological momentary assessments for the detection of prolonged stress in real life. *bioRxiv*.

Umematsu, T., Sano, A., Taylor, S., Tsujikawa, M., & Picard, R. W. (2020). Forecasting stress, mood, and health from daytime physiology in office workers and students. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, *2020-July*, 5953–5957. https://doi.org/10.1109/EMBC44109.2020.9176706

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE*, *14*(11), 1–20. https://doi.org/10.1371/journal.pone.0224365

van der Zwan, J. E., de Vente, W., Huizink, A. C., Bögels, S. M., & de Bruin, E. I. (2015). Physical Activity, Mindfulness Meditation, or Heart Rate Variability Biofeedback for Stress Reduction: A Randomized Controlled Trial. *Applied Psychophysiology Biofeedback*. https://doi.org/10.1007/s10484-015-9293-x

Van Diest, I., Verstappen, K., Aubert, A. E., Widjaja, D., Vansteenwegen, D., & Vlemincx, E. (2014). Inhalation/Exhalation Ratio Modulates the Effect of Slow Breathing on Heart Rate Variability and Relaxation. *Applied Psychophysiology Biofeedback*, *39*(3-4), 171–180. https://doi.org/10.1007/s10484-014-9253-x

Vaschillo, E., Lehrer, P., Rishe, N., & Konstantinov, M. (2002). Heart rate variability biofeedback as a method for assessing baroreflex function: A preliminary study of resonance in the cardiovascular system. *Applied Psychophysiology Biofeedback*, *27*(1), 1–27. https://doi.org/10.1023/A:1014587304314

Vaschillo, E. G., Vaschillo, B., & Lehrer, P. M. (2006). Characteristics of resonance in heart rate variability stimulated by biofeedback. *Applied Psychophysiology Biofeedback*. https://doi.org/10.1007/s10484-006-9009-3

Vollmer, M., Bläsing, D., & Kaderali, L. (2019). Alignment of Multi-Sensored Data: Adjustment of Sampling Frequencies and Time Shifts. *2019 Computing in Cardiology Conference (CinC)*, *45*, 2019–2022. https://doi.org/10.22489/cinc.2019.031

Volterrani, M., Scalvini, S., Mazzuero, G., Lanfranchi, P., Colombo, R., Clark, A. L., & Levi, G. (1994). Decreased heart rate variability in patients with chronic obstructive pulmonary disease. *Chest*, *106*(5), 1432–1437. https://doi.org/10.1378/chest.106.5.1432

Voss, A., Schroeder, R., Heitmann, A., Peters, A., & Perz, S. (2015). Short-term heart rate variability - Influence of gender and age in healthy subjects. *PLoS ONE*, *10*(3), 1–33. https://doi.org/10.1371/journal.pone.0118308

Vrijkotte, T. G., Van Doornen, L. J., & De Geus, E. J. (2000). Effects of Work Stress on Ambulatory Blood Pressure, Heart Rate, and Heart Rate Variability. *Hypertension*, *35*(4), 880–886. https://doi.org/10.1161/01.HYP.35.4.880

Watson, D., & Clark, L. A. (1994). THE PANAS-X Manual for the Positive and Negative Affect Schedule - Expanded Form.

Weerdmeester, J., Van Rooij, M. M., Engels, R. C., & Granic, I. (2020). An Integrative Model for the Effectiveness of Biofeedback Interventions for Anxiety Regulation: Viewpoint. *Journal of Medical Internet Research*, *22*(7). https://doi.org/10.2196/14958

Wehrwein, E. A., Orer, H. S., & Barman, S. M. (2016). Overview of the Anatomy, Physiology, and Pharmacology of the Autonomic Nervous System. *Comprehensive Physiology*, *6*(3), 1239–1278. https://doi.org/10.1002/CPHY.C150037

Wells, R., Outhred, T., Heathers, J. A., Quintana, D. S., & Kemp, A. H. (2012). Matter Over Mind: A Randomised-Controlled Trial of Single-Session Biofeedback Training on Performance Anxiety and Heart Rate Variability in Musicians. *PLoS ONE*, *7*(10). https://doi.org/10.1371/journal.pone.0046597

WHO. (2001). *The World Health Report 2001: Mental Health: New Understanding, New Hope*. World Health Organization.

WHO. (2004). *Promoting Mental Health; Concepts emerging evidence and practice. Summary Report* (tech. rep.). World Health Organisation. Geneva.

WHO. (2022a). *Mental disorders*. URL: https://www.who.int/news-room/fact-sheets/detail/mental-disorders

WHO. (2022b). *Mental Health and COVID-19 : Early evidence of the pandemic's impact* (tech. rep. March).

Williams, D. W. P., Thayer, J. F., & Koenig, J. (2016). Resting cardiac vagal tone predicts intraindividual reaction time variability during an attention task in a sample of young and healthy adults. *Psychophysiology*, *53*(12), 1843–1851. https://doi.org/10.1111/psyp.12739

Won, E., & Kim, Y.-K. (2016). Stress, the Autonomic Nervous System, and the Immune-kynurenine Pathway in the Etiology of Depression. *Current Neuropharmacology*, *14*(7), 665–673. https://doi.org/10.2174/1570159x14666151208113006

Wu, L., Shi, P., Yu, H., & Liu, Y. (2020). An optimization study of the ultra-short period for HRV analysis at rest and post-exercise. *Journal of Electrocardiology*, *63*, 57–63. https://doi.org/10.1016/j.jelectrocard.2020.10.002

Wylie, S. A., van den Wildenberg, W. P., Ridderinkhof, K. R., Bashore, T. R., Powell, V. D., Manning, C. A., & Wooten, G. F. (2009). The effect of speed-accuracy strategy on response interference control in Parkinson's disease. *Neuropsychologia*, *47*(8-9), 1844–1853. https://doi.org/10.1016/j.neuropsychologia.2009.02.025

Yan, Y., Zhang, J. W., Zang, G. Y., & Pu, J. (2019). The primary use of artificial intelligence in cardiovascular diseases: What kind of potential role does artificial intelligence play in future medicine? *Journal of Geriatric Cardiology*, *16*(8), 585–591. https://doi.org/10.11909/j.issn.1671-5411.2019.08.010

Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, *1168*(2). https://doi.org/10.1088/1742-6596/1168/2/022022

You, M., Laborde, S., Salvotti, C., Zammit, N., Mosley, E., & Dosseville, F. (2021a). Influence of a Single Slow-Paced Breathing Session on Cardiac Vagal Activity in Athletes. *International Journal of Mental Health and Addiction*. https://doi.org/10.1007/s11469-020-00467-x

You, M., Laborde, S., Zammit, N., Iskra, M., Borges, U., Dosseville, F., & Vaughan, R. S. (2021b). Emotional intelligence training: Influence of a brief slow-paced breathing exercise on psychophysiological variables linked to emotion regulation. *International Journal of Environmental Research and Public Health*, *18*(12). https://doi.org/10.3390/ijerph18126630

Yu, B., An, P., Hendriks, S., Zhang, N., Feijs, L., Li, M., & Hu, J. (2021). ViBreathe: Heart Rate Variability Enhanced Respiration Training for Workaday Stress Management via an Eyes-free Tangible Interface. *International Journal of Human–Computer Interaction*, *00*(00), 1–20. https://doi.org/10.1080/10447318.2021.1898827

Yu, B., Funk, M., Hu, J., & Feijs, L. (2018a). Unwind: a musical biofeedback for relaxation assistance. *Behaviour and Information Technology*, *37*(8), 800–814. https://doi.org/10.1080/0144929X.2018.1484515

Yu, B., Funk, M., Hu, J., Wang, Q., & Feijs, L. (2018b). Biofeedback for everyday stress management: A systematic review. *Frontiers in ICT*, *5*. https://doi.org/10.3389/fict.2018.00023

Zaccaro, A., Piarulli, A., Laurino, M., Garbella, E., Menicucci, D., Neri, B., & Gemignani, A. (2018). How Breath-Control Can Change Your Life: A Systematic Review on Psycho-Physiological Correlates of Slow Breathing. *Frontiers in Human Neuroscience*, *12*, 1–16. https://doi.org/10.3389/fnhum.2018.00353

Zhai, J., & Barreto, A. (2006). Stress recognition using non-invasive technology. *FLAIRS 2006 - Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference*, *2006*, 395–401.

Zhang, H., Zhu, M., Zheng, Y., & Li, G. (2015). Toward capturing momentary changes of heart rate variability by a dynamic analysis method. *PLoS ONE*, *10*(7), 1–13. https://doi.org/10.1371/journal.pone.0133148

Zheng, W. L., & Lu, B. L. (2016). Personalizing EEG-based affective models with transfer learning. *IJCAI International Joint Conference on Artificial Intelligence*, *2016-Janua*, 2732–2738.

Appendix A

Questionnaires

A.1 Demographic

1. What is your gender?
2. What is your age?
3. Weight (Kg): Height (cm):
4. What is your current occupation?

□ Student □ PT Employed □ FT Employed □ Retired □ Unemployed

5. What is the highest degree you have completed?

○ High School Diploma ○ Bachelor's degree ○ Postgraduate degree
○ Professional degree ○ Other, please specify

6. What is your experience level with the following items:

(1: No experience at all, 5: Expert)	1	2	3	4	5
Meditation/ Yoga exercises?	○	○	○	○	○
Deep breathing activities?	○	○	○	○	○

7. When did you get up this morning?
8. When did you go to sleep last night?
9. On a scale from 1 to 5, how would you rate the following about yourself?

(1: Poor/Low, 5: Excellent)	1	2	3	4	5
Your physical fitness	○	○	○	○	○
Your physical activity	○	○	○	○	○

A.2 HRV-Related Questionnaire

Please answer the following questions:

	Yes	No
1. Have you rushed in order to arrive on time for this experiment?	○	○
2. Have you consumed any caffeine beverages in the past two hours?	○	○
3. Have you consumed any alcoholic beverages in the past 24 hours?	○	○
4. Do you usually smoke? If yes, please report the number of cigarettes you smoke on a daily basis.	○	○
5. Have you smoked in the past two hours?	○	○
6. Have you eaten in the past two hours?	○	○
7. Do you suffer from any mental disorders, for example, severe depression or anxiety disorder?	○	○
8. Do you currently take any cardioactive medications such as anti-depressant or anti-hypertensive?	○	○
9. Do you have any chronic heart issues or respiratory conditions?	○	○
10. Do you have any known blood pressure conditions?	○	○
11. Did you follow your usual sleep routine last night?	○	○

A.3 GAD-7 | Generalised Anxiety Disorder Questionnaire

Over the last 2 weeks, how often have you been bothered by the following problems?

	1	2	3	4
1. Feeling nervous, anxious or on edge	○	○	○	○
2. Not being able to stop or control worrying	○	○	○	○
3. Worrying too much about different things	○	○	○	○
4. Trouble relaxing	○	○	○	○
5. Being so restless that it is hard to sit still	○	○	○	○
6. Becoming easily annoyed or irritable	○	○	○	○
7. Feeling afraid as if something awful might happen	○	○	○	○

A.4 DASS-21 | Depression, Anxiety and Stress Scale

Please read each statement and rate your answer. There are no right or wrong answers. Do not spend too much time on any statement.

0 – Did not apply to me at all
1 – Applied to me to some degree, or some of the time
2 – Applied to me to a considerable degree or a good part of time
3 – Applied to me very much or most of the time

	0	1	2	3
1. I found it hard to wind down	○	○	○	○
2. I was aware of dryness of my mouth	○	○	○	○
3. I couldn't seem to experience any positive feeling at all	○	○	○	○
4. I experienced breathing difficulty (e.g. excessively rapid breathing, breathlessness in the absence of physical exertion)	○	○	○	○
5. I found it difficult to work up the initiative to do things	○	○	○	○
6. I tended to over-react to situations	○	○	○	○
7. I experienced trembling (e.g. in the hands)	○	○	○	○
8. I felt that I was using a lot of nervous energy	○	○	○	○
9. I was worried about situations in which I might panic and make a fool of myself	○	○	○	○
10. I felt that I had nothing to look forward to	○	○	○	○
11. I found myself getting agitated	○	○	○	○
12. I found it difficult to relax	○	○	○	○
13. I felt down-hearted and blue	○	○	○	○
14. I was intolerant of anything that kept me from getting on with what I was doing	○	○	○	○
15. I felt I was close to panic	○	○	○	○
16. I was unable to become enthusiastic about anything	○	○	○	○
17. I felt I wasn't worth much as a person	○	○	○	○
18. I felt that I was rather touchy	○	○	○	○
19. I was aware of the action of my heart in the absence of physical exertion (e.g. sense of heart rate increase, heart missing a beat)	○	○	○	○
20. I felt scared without any good reason	○	○	○	○
21. I felt that life was meaningless	○	○	○	○

A.5 PANAS | Positive and Negative Affect Schedule

This scale consists of a number of words and phrases that describe different feelings and emotions. Indicate to what extent you have felt this way during the last week.

	very slightly or not at all	a little	moderately	quite a bit	extremely
1. attentive	○	○	○	○	○
2. strong	○	○	○	○	○
3. irritable	○	○	○	○	○
4. inspired	○	○	○	○	○
5. afraid	○	○	○	○	○
6. alert	○	○	○	○	○
7. upset	○	○	○	○	○
8. active	○	○	○	○	○
9. guilty	○	○	○	○	○
10. nervous	○	○	○	○	○
11. excited	○	○	○	○	○
12. hostile	○	○	○	○	○
13. proud	○	○	○	○	○
14. jittery	○	○	○	○	○
15. ashamed	○	○	○	○	○
16. scared	○	○	○	○	○
17. enthusiastic	○	○	○	○	○
18. distressed	○	○	○	○	○
19. determined	○	○	○	○	○
20. interested	○	○	○	○	○

A.6 IPAQ | International Physical Activity Questionnaire

We are interested in finding out about the kinds of physical activities that people do as part of their everyday lives. The questions will ask you about the time you spent being physically active in the **last 7 days**. Please answer each question even if you do not consider yourself to be an active person. Please think about the activities you do at work, as part of your house and yard work, to get from place to place, and in your spare time for recreation, exercise or sport.

Vigorous Activities

Think about all the **vigorous** activities that you did in the last 7 days. Vigorous physical activities refer to activities that take hard physical effort and make you breathe much harder than normal. Think only about those physical activities that you did for at least 10 minutes at a time.

1. During the **last 7 days**, on how many days did you do vigorous physical activities like heavy lifting, digging, aerobics, or fast bicycling?
   _____ days per week
   □ No vigorous physical activities ⟶ skip to question 3

2. How much time did you usually spend doing **vigorous** physical activities on one of those days?
   _____ hours per day
   _____ minutes per day
   □ Don't know/Not sure

Moderate Activities

Think about all the **moderate** activities that you did in the last 7 days. Moderate activities refer to activities that take moderate physical effort and make you breathe somewhat harder than normal. Think only about those physical activities that you did for at least 10 minutes at a time.

3. During the **last 7 days**, on how many days did you do moderate physical activities like carrying light loads, bicycling at a regular pace, or doubles tennis? Do not include walking.
   _____ days per week
   □ No moderate physical activities ⟶ skip to question 5

4. How much time did you usually spend doing **moderate** physical activities on one of those days?
   _____ hours per day
   _____ minutes per day
   □ Don't know/Not sure

Walking

Think about the time you spent **walking** in the last 7 days. This includes at work and at home, walking to travel from place to place, and any other walking that you have done solely for recreation, sport, exercise, or leisure.

5. During the **last 7 days**, on how many days did you walk for at least 10 minutes at a time?
   ______ days per week
   □ No walking ⟶ skip to question 7

6. How much time did you usually spend **walking** on one of those days?
   ______ hours per day
   ______ minutes per day
   □ Don’t know/Not sure

Sitting

The last question is about the time you spent **sitting** on weekdays during the last 7 days. Include time spent at work, at home, while doing course work and during leisure time. This may include time spent sitting at a desk, visiting friends, reading, or sitting or lying down to watch television.

7. During the **last 7 days**, how much time did you spend **sitting** on a **week day**?
   ______ hours per day
   ______ minutes per day
   □ Don’t know/Not sure

A.7 PSQI | Pittsburgh Sleep Quality Index

The following questions relate to your usual sleep habits during the past month only. Your answers should indicate the most accurate reply for the majority of days and nights in the **past month**. Please answer all questions.

1. During the past month, what time have you usually gone to bed at night?
   BED TIME: ______
2. During the past month, how long (in minutes) has it usually taken you to fall asleep each night?
   NUMBER OF MINUTES: ______
3. During the past month, what time have you usually gotten up in the morning?
   GETTING UP TIME: ______
4. During the past month, how many hours of actual sleep did you get at night? (This may be different than the number of hours you spent in bed.)
   HOURS OF SLEEP PER NIGHT: ______

For each of the remaining questions, check the one best response.

0 – Not during the past month

1 – Less than once a week

2 – Once or twice a week

3 – Three or more times a week

5. During the past month, how often have you had trouble sleeping because you . . .

	0	1	2	3
a. Cannot get to sleep within 30 minutes	○	○	○	○
b. Wake up in the middle of the night or early morning	○	○	○	○
c. Have to get up to use the bathroom	○	○	○	○
d. Cannot breathe comfortably	○	○	○	○
e. Cough or snore loudly	○	○	○	○
f. Feel too cold	○	○	○	○
g. Feel too hot	○	○	○	○
h. Had bad dreams	○	○	○	○
i. Have pain	○	○	○	○

During the past month ...	0	1	2	3
6. how often have you taken medicine to help you sleep (prescribed or "over the counter")?	○	○	○	○
7. how often have you had trouble staying awake while driving, eating meals, or engaging in social activity?	○	○	○	○

8. During the past month, how would you rate your sleep quality overall?

   ○ Very good
   ○ Fairly good
   ○ Fairly bad
   ○ Very bad

9. During the past month, how much of a problem has it been for you to keep up enough enthusiasm to get things done?

   ○ No problem at all
   ○ Only a very slight problem
   ○ Somewhat of a problem
   ○ A very big problem

A.8 Affective State Questionnaire

1. How stressed you are right now? (1: Not Stressed at all, 5: Extremely Stressed)

1	2	3	4	5
○	○	○	○	○

2. How are you feeling right now? (1: Very Sad, 5:Very Happy)

1	2	3	4	5
○	○	○	○	○

3. This scale consists of a number of words and phrases that describe different feelings and emotions. Read each item and then mark the appropriate answer in the space next to that word. Indicate to what extent you have felt this way at this moment. Use the following scale to record your answers:

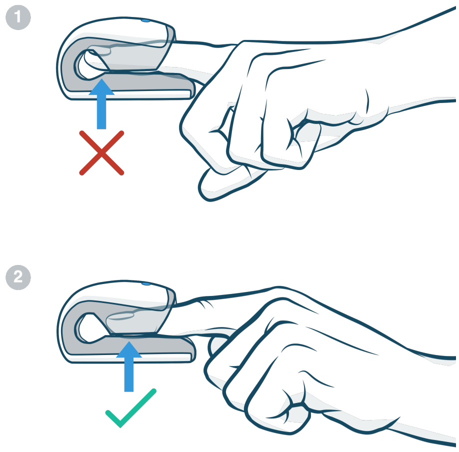
	Very slightly or not at all	A little	Moderately	Quite a bit	Extremely
1. calm	○	○	○	○	○
2. attentive	○	○	○	○	○
3. sluggish	○	○	○	○	○
4. relaxed	○	○	○	○	○
5. tired	○	○	○	○	○
6. sleepy	○	○	○	○	○
7. at ease	○	○	○	○	○
8. drowsy	○	○	○	○	○
9. determined	○	○	○	○	○
10. concentrating	○	○	○	○	○

Appendix B

Heart Rate Variability

B.1 CorSense

(a) CorSense



(b) Finger Attachment Instructions

(c) EliteHRV Application

Figure B.1: CorSense by EliteHRV

B.2 Dataset Review

Table B.1: Applicability Review for the Public Datasets

Dataset Name	Reference	Applicable?	Description
MIT-BIH Normal	Goldberger et al., 2000	No	Long term recordings (i.e., not a controlled experiment; thus, activity condition was not specified)
Yoga and Chi	Peng et al., 1999	No	Study involved controlled breathing activity
Normal Group	Peng et al., 1999	Yes	Chapter 4
Ironman	Peng et al., 1999	No	Participants were athletes.
WESAD	Schmidt et al., 2018	Yes	Chapter 7
SWELL	Koldijk et al., 2014	Yes	Chapter 7
Spiders	Ihmig et al., 2020	No	Stress inducer: video clips (i.e., not a mental stress task)
BigIdeasLab	Bent et al., 2020	No	Study involved a physical activity
DEAP	Koelstra et al., 2012	No	Stress inducer: video clips (i.e., not a mental stress task)
SPM	Vollmer et al., 2019	No	Baseline HRV was collected during a standing position (i.e., not seated) and the study involved a physical activity
DaLia	Reiss et al., 2019	No	Study involved a physical activity
CogLoad	Gjoreski et al., 2020	No	No HRV data for baseline condition

B.3 GATT Heart Rate Measurement

Description

The Heart Rate Measurement characteristic is a variable-length structure containing a Flags field, a Heart Rate Measurement Value field and, based on the contents of the Flags field, may contain additional fields such as Energy Expended or RR-Interval.

Flags Field

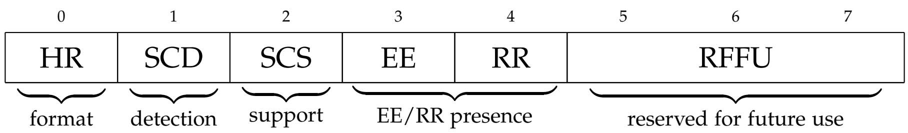


Figure B.2: Heart Rate Service Flag Byte.

Table B.2: Detailed Description of the Flag Byte

Bit	Acronym	Definition	Value
0	HR	Heart Rate Value Format	`0:UINT8, 1:UNIT16`
1	SCD	Sensor contact detected	`0:false, 1:true`
2	SCS	Sensor contact supported	`0:false, 1:true`
3	EE	Energy Expended present	`0:false, 1:true`
4	RR	RR intervals present	`0:false, 1:true`
5-7	RFU	Reserved for Future Use	

Appendix C

CH4 | Supplementary Materials

List of Tables

List of Figures

Table C.1: Mean of HRV measures obtained from each correction method compared with **KUB-corrected signal** (benchmark). Data are presented as mean ± standard deviation (mean absolute error [%])

Feature	KUB-corrected	Deletion	Window Average	Cubic Spline	Original	Erroneous
Time-Domain						
MeanRR	950.93 ± 169.26	953.71 ± 171.39 (0.3)	952.62 ± 170.13 (0.2)	952.76 ± 170.12 (0.2)	948.62 ± 166.6 (-0.2)	983.26 ± 169.42 (3.4)
SDNN	65.16 ± 12.32	65.01 ± 12.44 (-0.2)	65.06 ± 12.37 (-0.2)	65.05 ± 12.38 (-0.2)	65.28 ± 12.28 (0.2)	62.92 ± 11.7 (-3.4)
RMSSD	66.05 ± 30.02	76.94 ± 51.56 (16.5)	76.98 ± 50.05 (16.5)	81.83 ± 49.29 (23.9)	64.04 ± 39.36 (-3)	210.66 ± 24.93 (218.9)
NN50	130.16 ± 80.46	127.91 ± 88.7 (-1.7)	131.8 ± 91.31 (1.3)	147.95 ± 86.02 (13.7)	116.27 ± 81.54 (-10.7)	174.41 ± 69.57 (34)
pNN50	30.09 ± 21.54	31.33 ± 23.98 (4.1)	30.58 ± 23.68 (1.6)	33.97 ± 22.64 (12.9)	29.62 ± 23.36 (-1.6)	40.44 ± 19.59 (34.4)
Frequency-Domain						
VLF Power	442.06 ± 655.52	523.49 ± 719.06 (18.4)	453.81 ± 639.25 (2.7)	410.99 ± 583.9 (-7)	514.03 ± 882.67 (16.3)	480.88 ± 472.3 (8.8)
LF Power	2308.56 ± 2370.28	2900.82 ± 3090.94 (25.7)	2767.12 ± 3021.2 (19.9)	2606.03 ± 3035.99 (12.9)	2543.99 ± 2771.45 (10.2)	10011.8 ± 3730.3 (333.7)
HF Power	1778.14 ± 1773.71	2853.45 ± 3571.53 (60.5)	2709.2 ± 3373.93 (52.4)	2783.07 ± 3369.31 (56.5)	2180.36 ± 2713.63 (22.6)	14600.27 ± 3621.34 (721.1)
LFnu	55.18 ± 14.05	57.39 ± 15.88 (4)	56.66 ± 16.52 (2.7)	50.47 ± 13.23 (-8.5)	58.9 ± 17.51 (6.7)	40.14 ± 6.23 (-27.3)
HFnu	44.75 ± 14.04	42.56 ± 15.87 (-4.9)	43.28 ± 16.52 (-3.3)	49.45 ± 13.23 (10.5)	41.07 ± 17.52 (-8.2)	59.77 ± 6.2 (33.6)
Total Power	1.48 ± 0.88	1.83 ± 1.46 (23.6)	1.75 ± 1.34 (18.2)	1.18 ± 0.63 (-20.3)	2.05 ± 1.68 (38.5)	0.69 ± 0.2 (-53.4)
Non-Linear						
SD1	1.19 ± 0.08	1.19 ± 0.08 (0)	1.24 ± 0.08 (4.2)	1.23 ± 0.07 (3.4)	1.18 ± 0.08 (-0.8)	1.04 ± 0.15 (-12.6)
SD2	1.54 ± 0.19	1.55 ± 0.19 (0.6)	1.58 ± 0.19 (2.6)	1.62 ± 0.18 (5.2)	1.57 ± 0.19 (1.9)	1.08 ± 0.32 (-29.9)
DFA1	1 ± 0.22	0.97 ± 0.26 (-3)	1.01 ± 0.23 (1)	0.98 ± 0.2 (-2)	1.01 ± 0.26 (1)	1.03 ± 0.13 (3)
DFA2	0.37 ± 0.14	0.39 ± 0.16 (5.4)	0.39 ± 0.16 (5.4)	0.38 ± 0.15 (2.7)	0.39 ± 0.15 (5.4)	0.22 ± 0.08 (-40.5)

Table C.2: Reliability of the correction methods on the HRV features compared to those obtained from the analysis of the **KUB-corrected signal** through the Intraclass Correlation Coefficients with their associated 95% confidence intervals

	Deletion		Window Average		Cubic Spline		Original		Erroneous	
Feature	ICC	95% CI	ICC	95% CI	ICC	95% CI	ICC	95% CI	ICC	95% CI
Time-Domain										
MeanRR	1	[1, 1]	1	[1, 1]	1	[1, 1]	1	[1, 1]	.98	[.23, 1]
SDNN	1	[1, 1]	1	[1, 1]	1	[1, 1]	1	[1, 1]	.98	[.33, .99]
RMSSD	.82	[.68, .89]	.82	[.68, .90]	.79	[.54, .89]	.93	[.88, .95]	.05	[-.01, .17]
NN50	.99	[.98, .99]	.99	[.98, .99]	.96	[.70, .99]	.98	[.64, .99]	.82	[.02, .94]
pNN50	.99	[.98, .99]	.99	[.98, .99]	.98	[.78, .99]	.99	[.99, 1.0]	.88	[.03, .97]
Frequency-Domain										
VLF Power	.96	[.92, .98]	.97	[.95, .98]	.96	[.94, .98]	.91	[.85, .94]	.89	[.83, .94]
LF Power	.86	[.76, .92]	.91	[.84, .95]	.91	[.85, .95]	.95	[.91, .97]	.17	[-.04, .45]
HF Power	.71	[.50, .83]	.72	[.54, .84]	.72	[.51, .83]	.88	[.79, .93]	.04	[-.02, .15]
Total Power	.36	[.13, .56]	.41	[.19, .60]	.64	[.42, .78]	.34	[.12, .54]	.11	[-.06, .30]
LFnu	.66	[.49, .78]	.73	[.59, .83]	.78	[.58, .87]	.74	[.59, .83]	.17	[-.05, .38]
HFnu	.66	[.5, .78]	.73	[.59, .83]	.78	[.58, .87]	.74	[.59, .83]	.17	[-.05, .39]
Non-Linear										
SD1	.92	[.87, .95]	.80	[.20, .92]	.79	[.33, .91]	.89	[.81, .93]	0	[-.12, .15]
SD2	.81	[.71, .88]	.74	[.60, .83]	.78	[.46, .89]	.84	[.75, .90]	.19	[-.07, .44]
DFA1	.87	[.79, .92]	.93	[.88, .96]	.94	[.89, .96]	.90	[.84, .94]	.71	[.56, .81]
DFA2	.96	[.92, .98]	.96	[.89, .98]	.97	[.95, .98]	.96	[.93, .98]	.30	[-.06, .57]

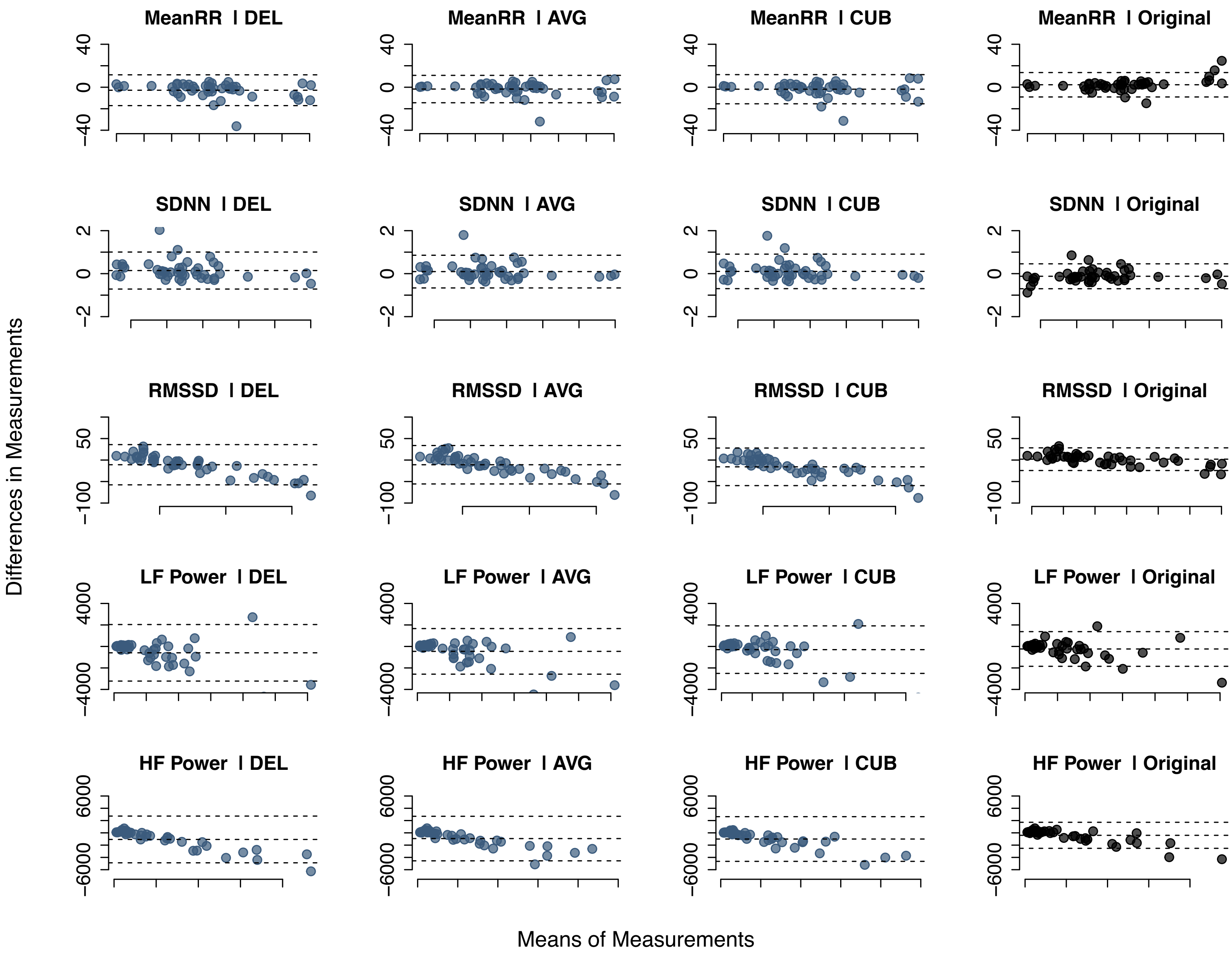


Figure C.1: Bland-Altman Plot with 95% Limits of Agreements Comparing HRV Measures of Various Correction Methods against KUB-Filtered Signal. *Note.* Results of Original Signal (artefact-free) is also Presented.

Appendix D

CH5 | Supplementary Materials

List of Figures

List of Tables

Queen Mary
University of London

Queen Mary, University of London
Room W104
Queen's Building
Queen Mary University of London
Mile End Road
London E1 4NS

Queen Mary Ethics of Research Committee
Hazel Covill
Research Ethics Facilitator
Tel: +44 (0) 20 7882 7915
Email: h.covill@qmul.ac.uk

c/o. Dr Tony Stockman
CS405
Department of Computer Science.
Queen Mary University of London
Mile End Road
London

9th December 2019

To Whom It May Concern:

Re: QMERC2019/58 – Investigation of multi-modal representation of HRV data for self-monitoring and analysis.

The above study was conditionally approved by The Queen Mary Ethics of Research Committee (Review Panel A) on the 4th September 2019; full approval was ratified by Chair's Action on the 14th November, with approval released by Facilitator (on receipt of paperwork) on the 9th December 2019.

This approval is valid for two years, (if the study is not started before this date then the applicant will have to reapply to the Committee).

Yours faithfully

Dr Helen Jenner – QMERC Chair.

Patron: Her Majesty the Queen
Incorporated by Royal Charter as Queen Mary and Westfield College, University of London

Figure D.1: Ethical Approval

Queen Mary
University of London

Consent form

Please complete this form after you have read the Information Sheet and/or listened to an explanation about the research.

Title of Study: Reducing mental stress using an audible representation of the heart rate variability

Queen Mary Ethics of Research Committee Ref: QMERC2019/58

Thank you for considering taking part in this research. The person organizing the research must explain the project to you before you agree to take part.

If you have any questions arising from the Information Sheet or explanation already given to you, please ask the researcher before you decide whether to join in. You will be given a copy of this Consent Form to keep and refer to at any time. If you are willing to participate in this study, please circle the appropriate responses and sign and date the declaration underneath.

Statement	Circle a response
I agree that the research project named above has been explained to me to my satisfaction in verbal and/or written form	YES / NO
I understand that if I decide at any other time during the research that I no longer wish to participate in this project, I can notify the researchers involved and be withdrawn from it immediately	YES / NO
I have read both the notes written above and the Information Sheet about the project, and understand what the research study involves	YES / NO
I agree to take part in the study, which will include use of my personal data	YES / NO

Participant's Signature:__________________________ Date: ___________________

Investigator's Statement:

I, Mariam, confirm that I have carefully explained the nature, demands and any foreseeable risks (where applicable) of the proposed research to the volunteer and provided a copy of this form

Figure D.2: Consent Form

Table D.1: Summary Statistics of the Log-Transformed HRV Measures in the **5-min Recording**

Feature	Baseline		Stress		Breathing	
	M	SD	M	SD	M	SD
Time-Domain						
MeanRR	6.70	.12	6.59	.14	6.71	.15
RMSSD	3.69	.45	3.68	.60	3.91	.56
SDNN	4.20	.37	4.25	.45	4.53	.38
pNN50	2.65	.90	2.56	1.02	3.06	.82
Frequency-Domain						
VLF power	6.94	.73	6.81	1.08	6.93	1.25
LF power	7.28	.80	7.53	.78	8.47	.91
HF power	6.25	.85	6.44	1.21	6.49	1.02
LFnu	4.26	.20	4.27	.19	4.44	.17
HFnu	3.22	.47	3.18	.54	2.46	.61
LF/HF	1.04	.66	1.09	.73	1.98	.76
Total power	8.08	.68	8.20	.87	8.91	.98
Non-Linear						
SD1	3.35	.45	3.40	.56	3.67	.50
SD2	4.49	.36	4.54	.45	4.82	.37
SampEn	.30	.15	.16	.28	-.04	.26
DFA1	.24	.11	.27	.11	.31	.21
DFA2	-.23	.27	-.34	.17	-.68	.45

Table D.2: Summary Statistics of the Log-Transformed HRV Measures in the **120-s Segment**

Feature	Baseline		Stress		Breathing	
	M	SD	M	SD	M	SD
Time-Domain						
MeanRR	6.71	0.15	6.58	0.15	6.72	0.15
RMSSD	3.73	0.47	3.70	0.57	3.89	0.55
SDNN	4.16	0.40	4.19	0.44	4.49	0.38
pNN50	2.75	0.91	2.51	1.11	3.02	0.91
Frequency-Domain						
LF power	7.38	0.94	7.40	1.01	8.33	1.02
HF power	6.33	1.05	6.43	1.26	6.60	0.92
LFnu	4.27	0.18	4.23	0.30	4.40	0.15
HFnu	3.22	0.44	3.21	0.56	2.68	0.60
LF/HF	1.05	0.61	1.02	0.83	1.72	0.75
Total power	8.08	0.88	8.01	1.01	8.69	0.86
Non-Linear						
SD1	3.38	0.47	3.36	0.57	3.66	0.49
SD2	4.45	0.40	4.47	0.43	4.78	0.37

Table D.3: Summary Statistics of the Log-Transformed HRV Measures in the **60-s Segment**

Feature	Baseline		Stress		Breathing	
	M	SD	M	SD	M	SD
Time-Domain						
MeanRR	6.71	0.15	6.58	0.15	6.72	0.15
RMSSD	3.73	0.48	3.70	0.61	3.88	0.57
SDNN	4.16	0.41	4.16	0.48	4.49	0.37
pNN50	2.74	0.94	2.48	1.22	3.00	0.94
Frequency-Domain						
LF power	7.39	1.00	7.36	1.07	8.33	1.04
HF power	6.37	1.09	6.39	1.28	6.62	0.96
LFnu	4.26	0.19	4.23	0.27	4.40	0.16
HFnu	3.25	0.42	3.22	0.56	2.69	0.61
LF/HF	1.01	0.60	1.01	0.81	1.71	0.77
Total power	8.05	0.93	7.92	1.07	8.66	0.90
Non-Linear						
SD1	3.38	0.48	3.35	0.61	3.65	0.51
SD2	4.43	0.41	4.45	0.48	4.77	0.36

Table D.4: Summary Statistics of the Log-Transformed HRV Measures in the **30-s Segment**

Feature	Baseline		Stress		Breathing	
	M	SD	M	SD	M	SD
Time-Domain						
MeanRR	6.72	0.16	6.55	0.16	6.73	0.15
RMSSD	3.74	0.56	3.60	0.67	4.01	0.56
SDNN	4.09	0.45	4.06	0.58	4.55	0.38
pNN50	2.73	1.00	2.31	1.19	3.20	0.77
Frequency-Domain						
LF power	7.17	0.97	6.99	1.10	8.53	0.93
HF power	6.28	1.04	5.89	1.46	6.77	1.08
LFnu	4.18	0.37	4.23	0.26	4.39	0.18
HFnu	3.28	0.59	3.13	0.76	2.62	0.78
LF/HF	0.89	0.93	1.10	1.01	1.77	0.95
Total power	7.60	0.90	7.36	1.14	8.75	0.88

Table D.5: Summary Statistics of the Log-Transformed HRV Measures in the **20-s Segment**

Feature	Baseline		Stress		Breathing	
	M	SD	M	SD	M	SD
Time-Domain						
MeanRR	6.72	0.16	6.56	0.16	6.73	0.16
RMSSD	3.72	0.58	3.56	0.70	3.98	0.57
SDNN	4.09	0.47	4.00	0.55	4.50	0.39
pNN50	2.63	1.23	2.17	1.25	3.17	0.80
Frequency-Domain						
LF power	6.90	1.01	6.73	1.39	8.38	1.07
HF power	6.13	1.03	5.71	1.54	6.53	1.27
LFnu	4.12	0.43	4.16	0.40	4.38	0.23
HFnu	3.34	0.64	3.15	0.88	2.53	0.96
LF/HF	0.77	1.03	1.02	1.24	1.84	1.17
Total power	7.39	0.89	7.17	1.29	8.61	1.02

Table D.6: Summary Statistics of the Log-Transformed HRV Measures in the **10-s Segment**

Feature	Baseline		Stress		Breathing	
	M	SD	M	SD	M	SD
Time-Domain						
MeanRR	6.72	0.16	6.55	0.17	6.74	0.15
RMSSD	3.71	0.58	3.52	0.71	3.98	0.58
SDNN	3.98	0.43	3.87	0.53	4.50	0.40
pNN50	2.58	1.36	1.97	1.22	3.16	0.82
Frequency-Domain						
LF power	6.62	0.91	6.33	1.18	8.11	1.02
HF power	6.11	0.91	5.91	1.39	6.62	1.20
LFnu	3.98	0.59	3.95	0.55	4.29	0.33
HFnu	3.47	0.65	3.53	0.69	2.80	0.91
LF/HF	0.50	1.21	0.42	1.20	1.49	1.22
Total power	7.25	0.71	6.99	1.15	8.43	0.92

Table D.7: Bias [95% CI] calculated from Bland-Altman Analysis between 5-min and UST segments during **baseline** (%)

Feature	120	60	30	20	10
Time					
MeanRR	.44 [-11.3, 12.1]	.76 [-11, 12.5]	1.45 [-13.2, 16.1]	1.21 [-14, 16.4]	1.95 [-14.1, 18.1]
RMSSD	.97 [-23.6, 25.6]	3 [-35.7, 41.7]	2.21 [-43.3, 47.7]	1.78 [-46.4, 50]	1.4 [-51.6, 54.4]
SDNN	3.51 [-31.5, 38.5]	4.35 [-30.7, 39.4]	10.6 [-43.7, 64.9]	10.86 [-45.4, 67.1]	21.23 [-48.3, 90.8]
pNN50	10.08 [-22.8, 42.3]	9 [-34.3, 52.3]	7.38 [-54.3, 69.1]	2.52 [-121.5, 126.6]	8.82 [-155.5, 173.2]
Frequency					
LF power	8.53 [-95, 112.1]	14.75 [-112.7, 130.2]	7.73 [-163.2, 178.6]	28.38 [-149.5, 206.3]	50.38 [-113.5, 214.3]
HF power	8.56 [-81.8, 99]	10.02 [-85.5, 110]	2.32 [-116.9, 121.5]	11.34 [-98.2, 120.9]	12.15 [-115.1, 139.4]
LF/HF	1.05 [-94.4, 96.5]	1.84 [-109.2, 112.9]	12.93 [-145.5, 171.4]	22.43 [-148.7, 193.5]	38.76 [-137.8, 215.3]
Total power	.25 [-87.1, 87.6]	3.46 [-98.4, 105.3]	41.12 [-103.5, 185.7]	58.09 [-82.9, 199.1]	70.1 [-53.1, 193.3]
Non-Linear					
SD1	3.27 [-10.5, 17.1]	3.7 [-12.9, 20.3]			
SD2	4.63 [-34.4, 43.7]	5.74 [-34, 45.4]			

Table D.8: Bias [95% CI] calculated from Bland-Altman Analysis between 5-min and UST segments during **stress** (%)

Feature	120	60	30	20	10
Time					
MeanRR	1.85 [-4, 7.7]	1.7 [-4.3, 7.7]	4.15 [-7.6, 15.9]	3.62 [-9.6, 16.8]	4.73 [-9.9, 19.3]
RMSSD	4.33 [-31.5, 40.1]	4.78 [-35.3, 44.8]	11.76 [-42.2, 70.8]	12.24 [-45.9, 82.1]	15.05 [-51.9, 95.7]
SDNN	5.52 [-27, 38.1]	8.56 [-34.3, 51.4]	13.53 [-40.4, 77.5]	16.46 [-36.2, 85.1]	22.61 [-38.9, 110.1]
pNN50	4.71 [-56.8, 66.2]	14.11 [-107.4, 135.7]	17.64 [-98.5, 169.7]	18.58 [-109.5, 216.6]	20.65 [-96.4, 235.7]
Frequency					
LF power	10.02 [-80.4, 100.9]	13.59 [-86.7, 113.9]	41.94 [-108.6, 192.5]	52.95 [-107.3, 213.2]	87.17 [-58.7, 233]
HF power	0.36 [-101.3,102]	1.74 [-111.1,114.5]	46.28 [-109.5, 202.1]	58.54 [-97, 214.1]	41.31 [-130.5, 213.2]
LF/HF	6.17 [-97.7, 110]	6.53 [-109.1, 122.2]	1.5 [-144.4, 147.4]	6.4 [-167.2, 180]	52.33 [-126.1, 230.7]
Total power	14.17 [-76.5, 104.9]	22.22 [-73.1, 117.6]	64.45 [-73.3, 202.2]	73.46 [-66.9, 213.8]	92.21 [-32.2, 216.6]
Non-Linear					
SD1	4.32 [-31.5, 40.1]	4.8 [-35.3, 44.9]			
SD2	6.3 [-27.7, 40.3]	9.82 [-35, 54.6]			

Table D.9: Bias [95% CI] calculated from Bland-Altman Analysis between 5-min and UST segments during **paced breathing** (%)

Feature	120	60	30	20	10
Time					
MeanRR	.46 [-5.1, 6]	.44 [-5.8, 6.6]	2.13 [-5.7, 10]	2.16 [-6.8, 11.2]	2.5 [-6.4, 11.4]
RMSSD	.17 [-29.8, 30.1]	3.18 [-25.4, 31.7]	10.75 [-24.5, 44.4]	8.8 [-25.4, 43]	7.32 [-42.6, 57.3]
SDNN	4.3 [-27.4, 36]	4.41 [-27.3, 36.1]	2.06 [-41.5, 45.6]	2.99 [-51.6, 57.5]	3.02 [-59.4, 65.4]
pNN50	3.36 [-44, 50.7]	5.53 [-49.8, 60.9]	6.57 [-38.5, 65.7]	10.58 [-55.8, 77]	9.29 [-70.1, 88.6]
Frequency					
LF power	4.52 [-76.9, 104]	7.24 [-83, 109.5]	7.31 [-106.2, 120.8]	7.6 [-112.2, 127.4]	28.72 [-101.2, 158.6]
HF power	10.05 [-65.5, 85.8]	11.5 [-70.6, 93.6]	21.91 [-88, 131.8]	2.18 [-139.2, 143.6]	5.85 [-154.2, 165.9]
LF/HF	8.15 [-79.1, 125.4]	13.72 [-83.7, 131.2]	16.3 [-120.5, 153.1]	8.87 [-148.8, 166.6]	36.23 [-135.9, 208.4]
Total power	8.07 [-75.4, 111.5]	14.81 [-78.1, 119.8]	8.98 [-114.4, 132.4]	21.16 [-106.8, 149.1]	34.25 [-97.6, 166.1]
Non-Linear					
SD1	1.29 [-22.3, 24.9]	2.21 [-27.4, 31.9]			
SD2	4.69 [-28.3, 37.7]	4.96 [-27.6, 37.5]			

Table D.10: Pearson Correlation Coefficients of HRV Measures for the **Baseline** Condition with 95% CI

Feature	120	60	30	20	10
Time					
MeanRR	.99 [.98, .98]	.98 [.94, .97]	.95 [.86, .96]	.94 [.83, .96]	.92 [.80, .95]
RMSSD	.98 [.97,1.00]	.94 [.85, .99]	.92 [.74, .98]	.91 [.72, .97]	.89 [.70, .96]
SDNN	.90 [.82, .96]	.89 [.75, .96]	.78 [.52, .89]	.78 [.51, .91]	.59 [.19, .82]
pNN50	.98 [.96, .99]	.97 [.93, .99]	.95 [.87, .98]	.86 [.68, .95]	.82 [.59, .92]
Frequency					
LF power	.80 [.56, .92]	.74 [.45, .89]	.32 [-.15, .67]	.21 [-.26, .59]	.06 [-.39, .49]
HF power	.88 [.79, .95]	.87 [.76, .95]	.77 [.50, .90]	.81 [.57, .92]	.68 [.34, .86]
LF/HF	.68 [.34, .86]	.55 [.14, .80]	.35 [-.11, .69]	.29 [-.18, .65]	.36 [-.10, .69]
Total power	.85 [.65, .94]	.81 [.57, .92]	.44 [.00, .74]	.43 [-.01, .73]	.31 [-.16, .66]
Non-Linear					
SD1	.99 [.97, 1.00]	.98 [.96, .99]			
SD2	.86 [.78, .95]	.87 [.76, .95]			

Table D.11: Pearson Correlation Coefficients of HRV Measures for the **Stress** Condition with 95% CI

Feature	120	60	30	20	10
Time					
MeanRR	.98 [.96, .99]	.98 [.95, .99]	.93 [.82, .97]	.90 [.77, .97]	.90 [.76, .96]
RMSSD	.98 [.95, .99]	.97 [.93, .98]	.90 [.73, .96]	.86 [.71, .95]	.72 [.59, .83]
SDNN	.93 [.85, .97]	.89 [.76, .95]	.84 [.64, .94]	.81 [.57, .92]	.65 [.30, .85]
pNN50	.96 [.89, .98]	.89 [.73, .95]	.90 [.76, .96]	.84 [.63, .93]	.76 [.48, .90]
Frequency					
LF power	.85 [.66, .94]	.83 [.62, .90]	.51 [.09, .78]	.36 [-.10, .69]	.35 [-.11, .69]
HF power	.89 [.81, .95]	.85 [.77, .94]	.71 [.39, .88]	.71 [.39, .88]	.63 [.26, .84]
LF/HF	.76 [.47, .90]	.67 [.32, .86]	.54 [.13, .79]	.51 [.09, .78]	.27 [-.20, .64]
Total power	.79 [.54, .92]	.77 [.50, .91]	.51 [.09, .78]	.43 [-.01, .73]	.53 [.11, .79]
Non-Linear					
SD1	.95 [.87, .98]	.94 [.85, .98]			
SD2	.92 [.81, .97]	.87 [.78, .95]			

Table D.12: Pearson Correlation Coefficients of HRV Measures for the **Paced Breathing** Condition with 95% CI

Feature	120	60	30	20	10
Time					
MeanRR	.98 [.96, .99]	.98 [.95, .99]	.96 [.91, .99]	.95 [.89, .98]	.95 [.87, .97]
RMSSD	.96 [.91, .99]	.97 [.92, .98]	.95 [.88, .98]	.95 [.84, .97]	.90 [.75, .96]
SDNN	.95 [.81, .96]	.94 [.79, .96]	.86 [.76, .93]	.81 [.52, .88]	.74 [.48, .85]
pNN50	.96 [.91, .99]	.95 [.89, .98]	.94 [.85, .98]	.90 [.76, .96]	.86 [.67, .94]
Frequency					
LF power	.88 [.80, .95]	.87 [.76, .95]	.75 [.46, .89]	.76 [.47, .90]	.65 [.30, .85]
HF power	.92 [.81, .97]	.90 [.77, .96]	.77 [.50, .91]	.74 [.44, .89]	.54 [.13, .79]
LF/HF	.72 [.41, .88]	.68 [.35, .86]	.56 [.16, .81]	.55 [.14, .80]	.45 [.01, .74]
Total power	.78 [.52, .91]	.76 [.48, .90]	.56 [.15, .80]	.54 [.13, .80]	.39 [-.07, .71]
Non-Linear					
SD1	.97 [.93, .99]	.95 [.88, .98]			
SD2	.89 [.76, .96]	.89 [.75, .96]			

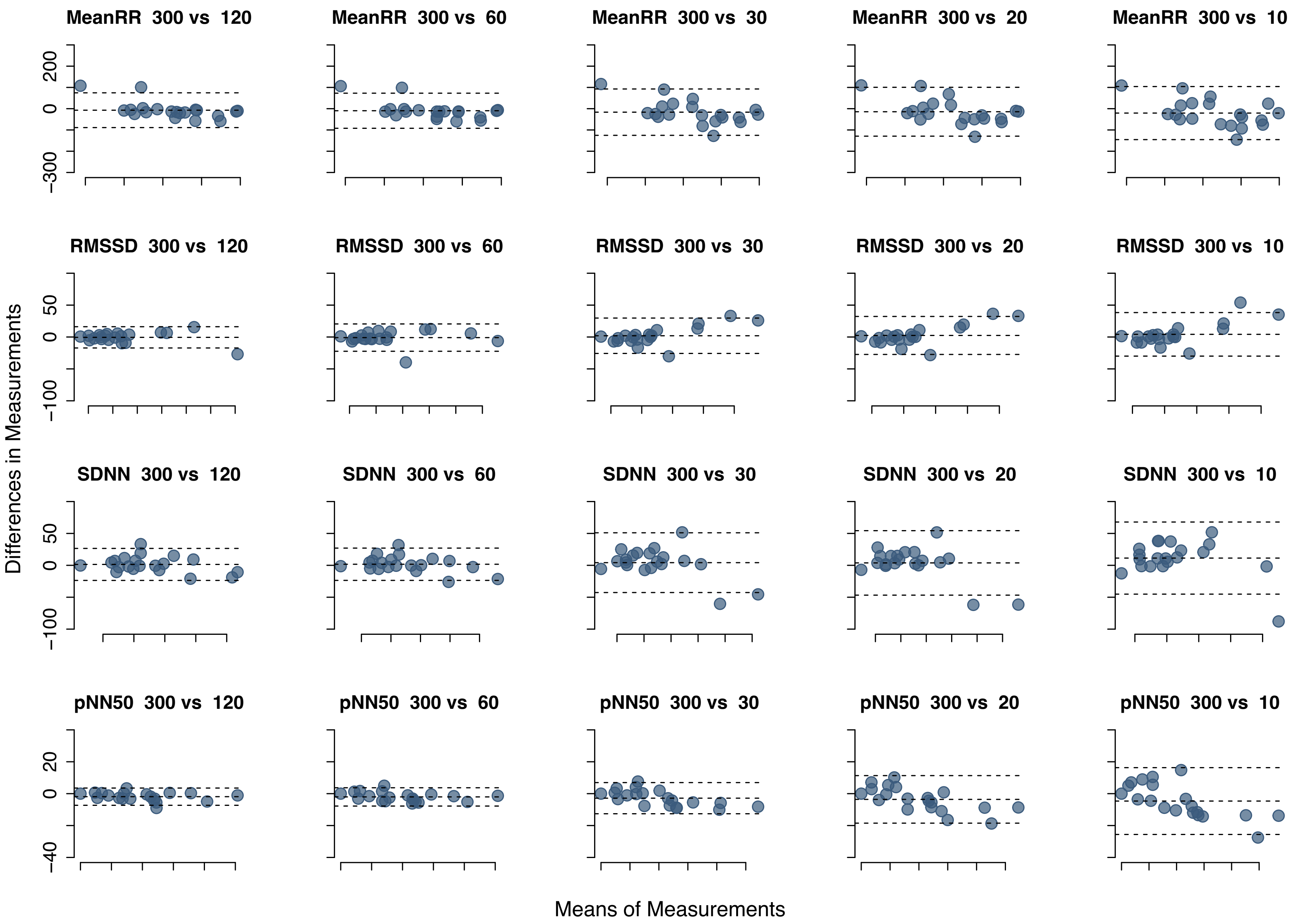


Figure D.3: Bland-Altman Plots with 95% Limits of Agreements of HRV **Time-Domain** Measures during **baseline**.

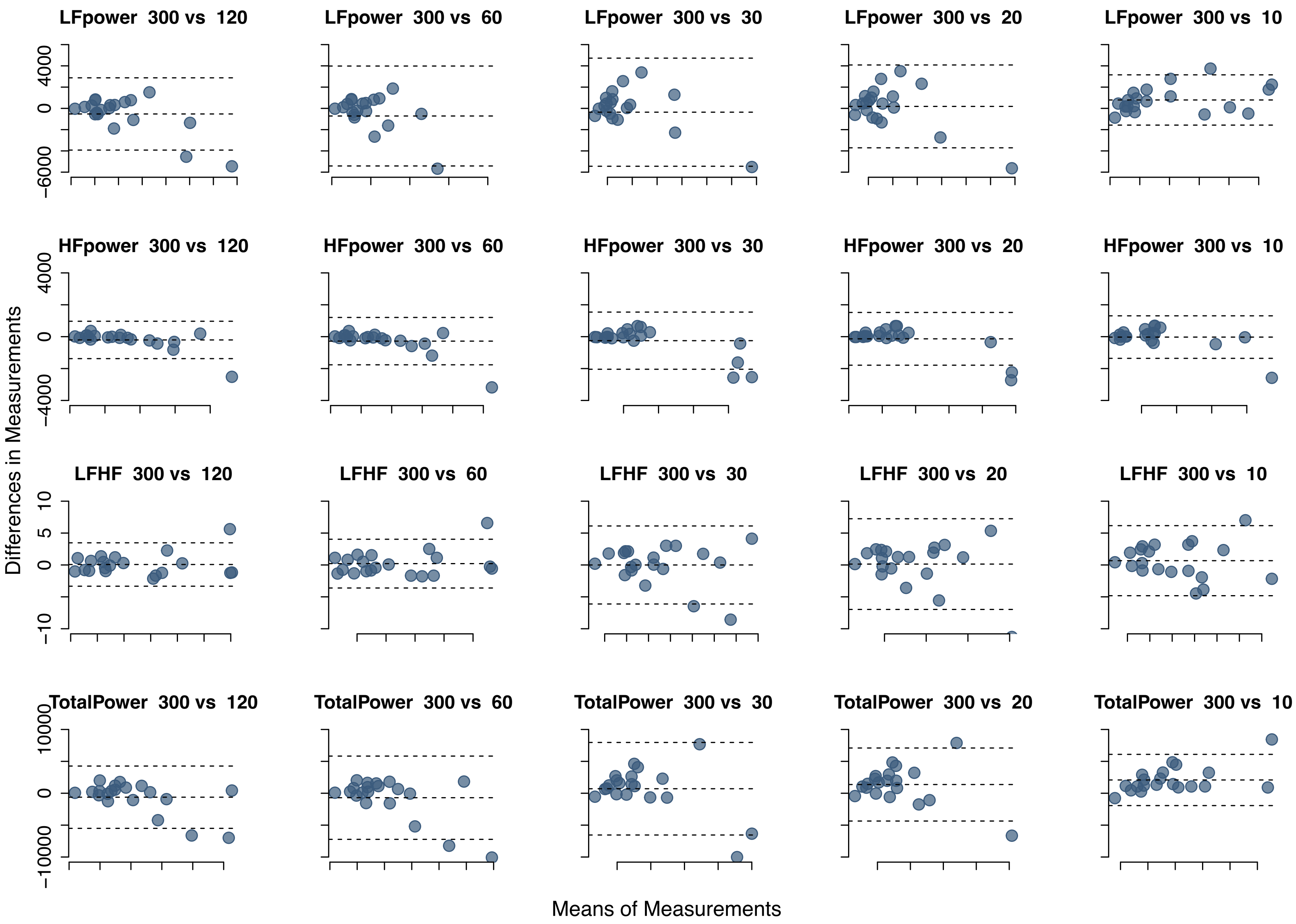


Figure D.4: Bland-Altman Plots with 95% Limits of Agreements of HRV Frequency-Domain Measures during **baseline**.

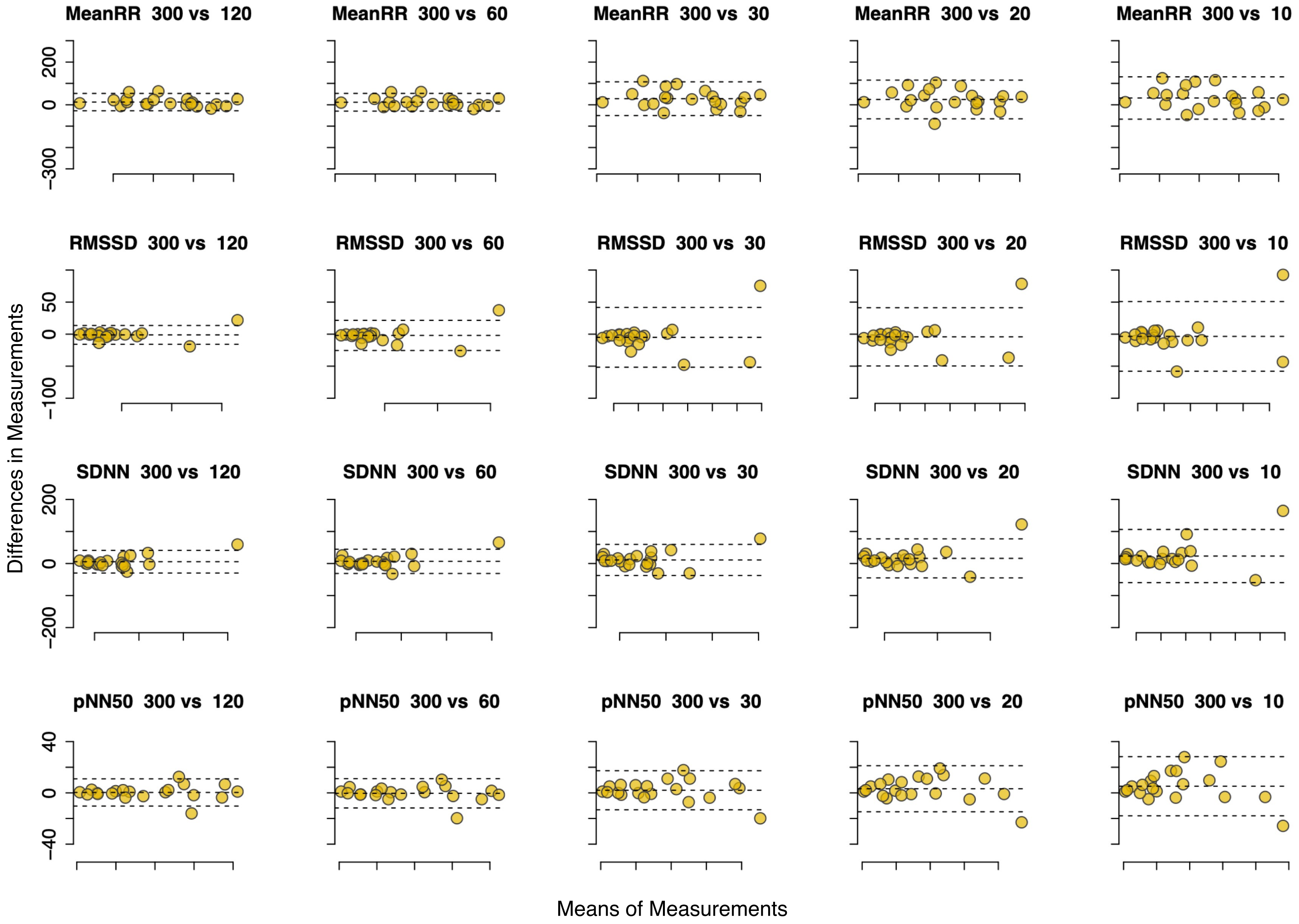


Figure D.5: Bland-Altman Plots with 95% Limits of Agreements of HRV Time-Domain Measures during **stress**.

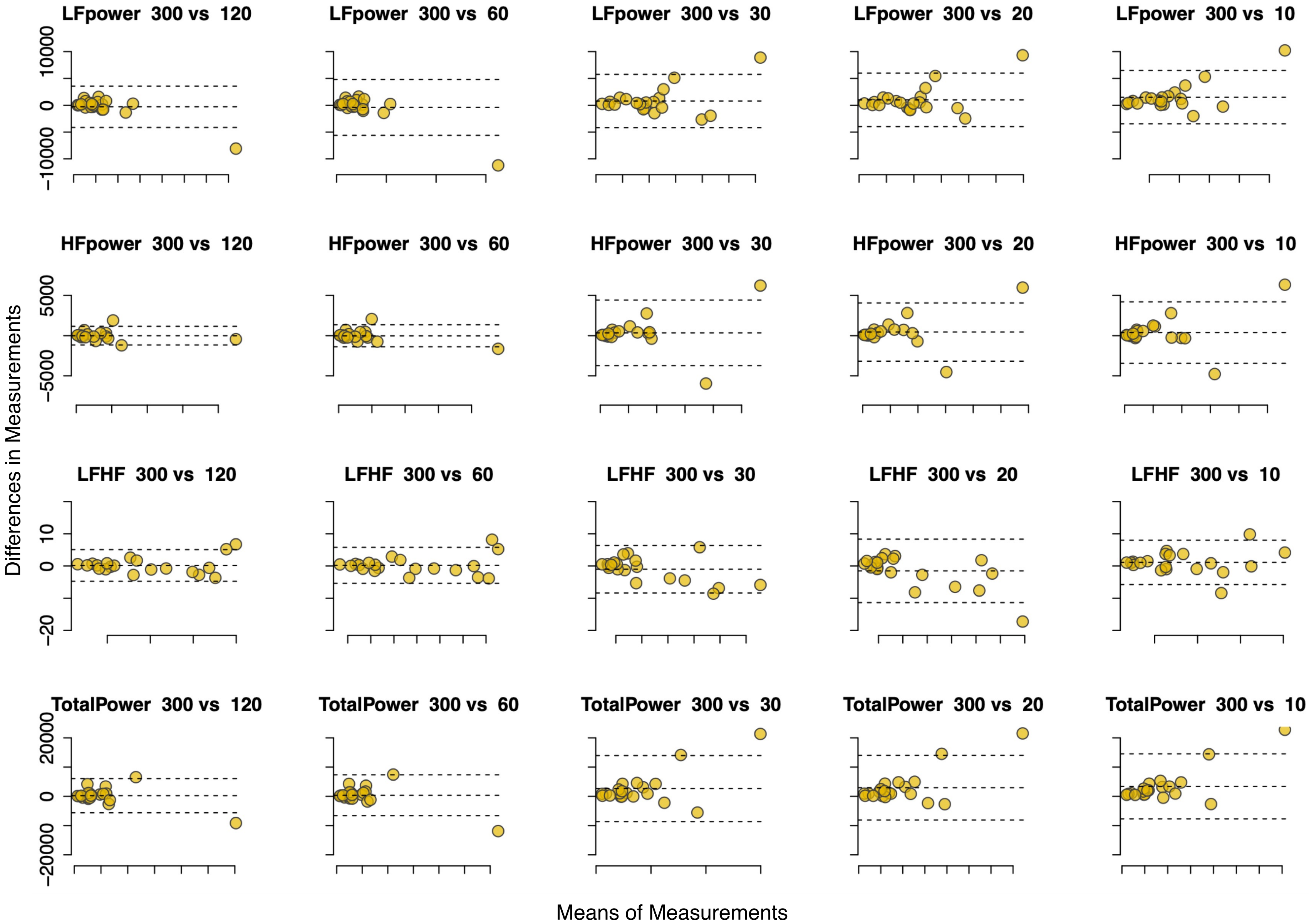


Figure D.6: Bland-Altman Plots with 95% Limits of Agreements of HRV Frequency-Domain Measures during **stress**.

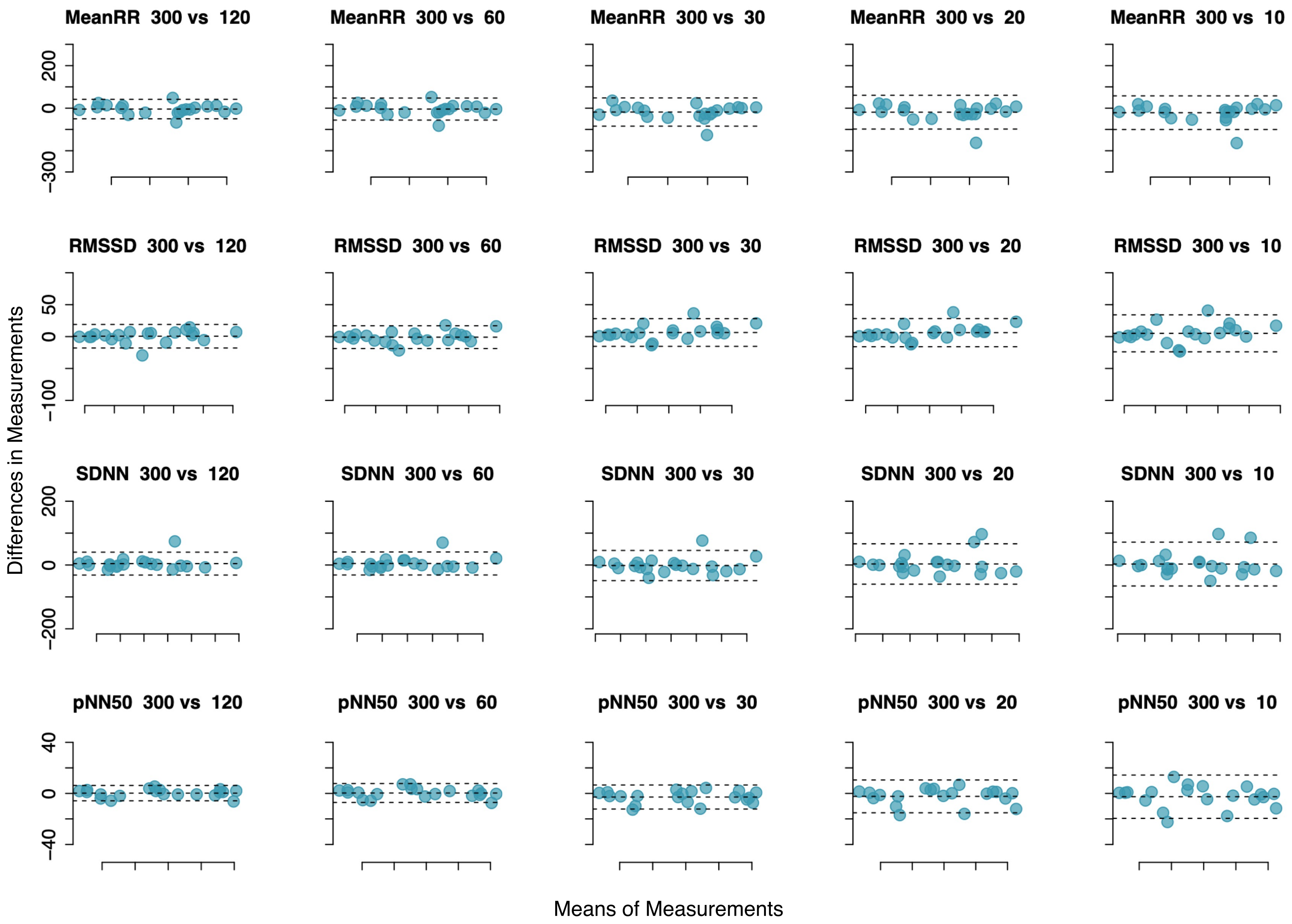


Figure D.7: Bland-Altman Plots with 95% Limits of Agreements of HRV Frequency-Domain Measures during **Paced Breathing**.

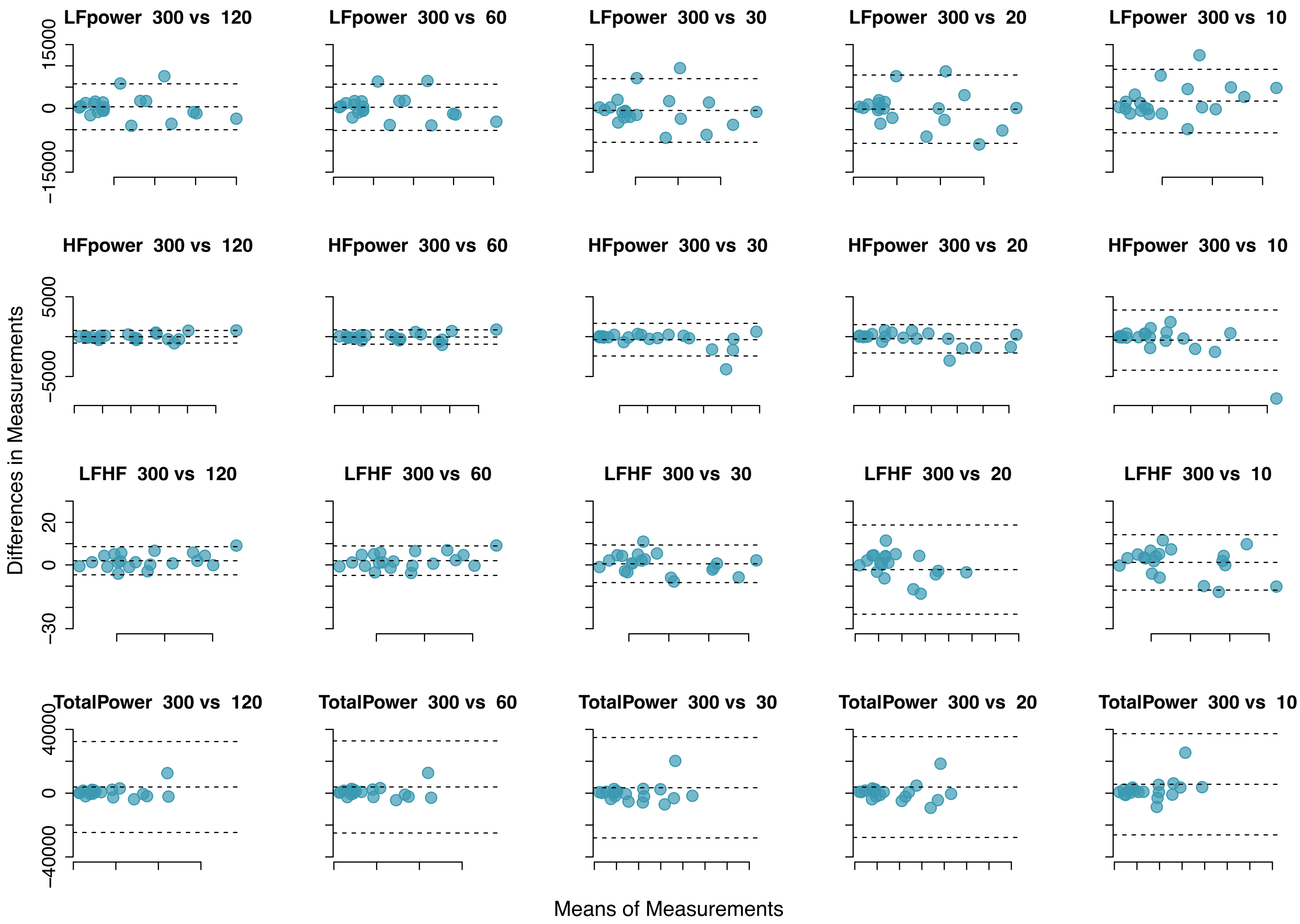


Figure D.8: Bland-Altman Plots with 95% Limits of Agreements of HRV Frequency-Domain Measures during **Paced Breathing**.

Table D.13: Multilevel Linear Model Analysis Results

Feature	5 min		2 min		1 min		30 sec		20 sec		10 sec	
	χ^2	p-value	χ^2	p-value	χ^2	p-value	χ^2	p-value	χ^2	p-value	χ^2	p-value
MeanRR	54.2	<.001	44.4	<.001	42.97	<.001	40.2	<.001	37.0	<.001	38.4	<.001
RMSSD	4.3	.117	2.6	.266	2.25	.324	8.3	.016	8.2	.016	9.2	.01
SDNN	12.0	.003	11.4	.003	12.63	.002	18.4	<.001	18.6	<.001	25.2	<.001
pNN50	7.6	.022	5.2	.074	5.08	.079	13.8	.001	13.0	.002	16.7	<.001
LF power	25.0	<.001	15.1	<.001	14.05	<.001	27.3	<.001	22.1	<.001	28.1	<.001
HF power	1.0	.601	1.0	.598	0.94	.626	8.2	.017	5.4	.069	5.0	.081
LF/HF	22.3	<.001	11.3	.003	11.57	.003	8.8	.013	9.2	.01	9.5	.009
Total power	13.3	.001	9.9	.007	9.72	.008	26.0	<.001	21.5	<.001	26.4	<.001
SD1	8.4	.015	7.1	.028	6.46	.04						
SD2	11.9	.003	11.9	.003	13.30	.001						

Table D.14: Detailed Summary of the Minimum Reliable UST Segment based on each Criterion

Feature	Analysis	Baseline	Stress	Paced Breathing
MeanRR	BA	10	10	10
	PC	10	10	10
	TA		10	10
RMSSD	BA	10	60	60
	PC	60	60	10
	TA		10	10
SDNN	BA	60	60	10
	PC	60	60	30
	TA		60	10
PNN50	BA	120	120	30
	PC	30	120	20
	TA		10	–
LF power	BA	120	120	20
	PC	–	–	60
	TA		120	10
HF power	BA	60	60	120
	PC	60	60	60
	TA		60	10
LF/HF	BA	60	20	120
	PC	–	–	–
	TA		–	10
Total power	BA	60	–	120
	PC	–	–	–
	TA		–	10
SD1	BA	60	60	60
	PC	60	60	60
	TA		–	10
SD2	BA	60	60	60
	PC	60	60	60
	TA		60	60

Note. BA is the Bland-Altman Analysis; where bias is less than 10%.
PC is the Pearson Correlation Analysis; where r > .80 95% CI lower bound > .75.
TA is the Trend Analysis for non-resting conditions.

Appendix E

CH6 | Supplementary Materials

List of Figures

List of Tables

List of Documents

QBRI
معهد قطر لبحوث الطب الحيوي
Qatar Biomedical Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Qatar Biomedical Research Institute
Institutional Review Board
IRB Approval Letter

Dr. Dena A. Al-Thani **Assistant Professor** College of Science and Engineering - Hamad Bin Khalifa University	
IRB Protocol Reference Number:	QBRI-IRB 2021-03-088
Project Title:	An Investigation into the impact of heart rate variability biofeedback on mental health state and cognitive performance
Review Type:	Expedited status
QBRI-IRB Approval Date:	25 March 2021
QBRI-IRB Expiration Date:	24 March 2022

The QBRI Institutional Review Board (IRB) has reviewed your application that was submitted the above referenced protocol (2021-03-088). It has been determined that your research proposal is eligible for **expedited status and Approved** for one year effective March 25 , 2021. This falls under the category (4) Collection of data through noninvasive procedures.
The research must be conducted according to the submitted research protocol outlined in the approved proposal. Please consider that any modifications to any aspect of the referenced study may render this approval invalid. Any amendments must be submitted to the IRB office and it cannot be implemented until the IRB approval has been given.

It is imperative that any serious adverse events experienced during the course of this study by research subjects, are immediately reported to the IRB office.

Request for a renewal, if required, should be submitted to the IRB at least 40 days prior to the expiry date to allow the IRB sufficient time to review and approve the request. It is the sole responsibility of the investigator to ensure the timely renewal of the IRB. Please note that it is the investigator's responsibility to ensure that they have a valid CITI certificate in the relevant curriculum during the course of the approval.

Wishing you all the success in conducting your research.

Dr. Khalid Al-Ali
Chairperson
KAlali

QATAR BIOMEDICAL RESEARCH INSTITUTE
IRB APPROVED
APPROVAL DATE: 25 MAR 2021
EXPIRATION DATE: 24 MAR 2022

Figure E.1: Ethical Approval

QATAR BIOMEDICAL RESEARCH INSTITUTE
IRB APPROVED
APPROVAL DATE: 25 MAR 2021
EXPIRATION DATE: 24 MAR 2022

QBRI
معهد قطر لبحوث الطب الحيوي
Qatar Biomedical Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

IRB number: .... Principal Investigator: Dr Dena Althani

Project Title: An Investigation into the impact of heart rate variability biofeedback on mental health state and cognitive performance

IC Version Date: ....

SIGNATURES

As a representative of this study, I have explained the purpose, the procedures, the possible benefits and risks that are involved in this research study. Any questions that have been raised have been answered to the individual's satisfaction.

Signature of Person Obtaining Consent **Date of Signature**

I, the undersigned have been informed about this study's purpose, procedures, possible benefits and risks, and I have received a copy of this consent. I have been given the opportunity to ask questions before I sign, and I have been told that I can ask other questions at any time. I voluntarily agree to be in this study. I am free to stop being in the study at any time without need to justify my decision and if I stop being in the study I understand it will not in any way affect my future treatment or medical management. I agree to cooperate with (name of principal investigator) and the research staff and to tell them immediately if I experience any unexpected or unusual symptoms.

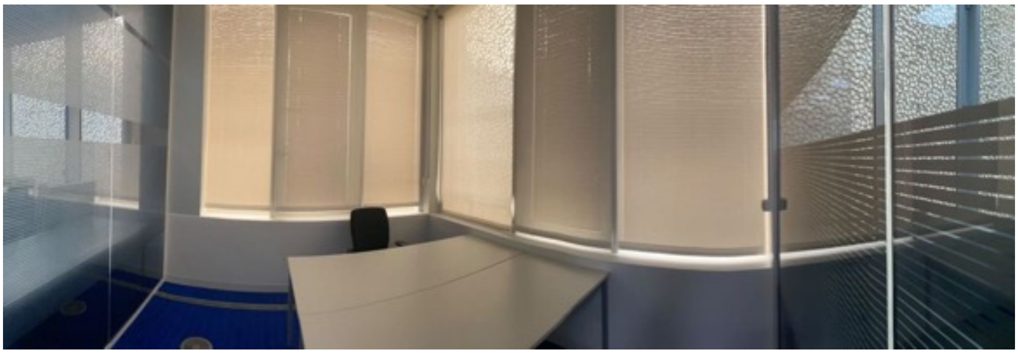
Participant's Signature **Date of Signature**

Signature of Witness **Date of Signature**

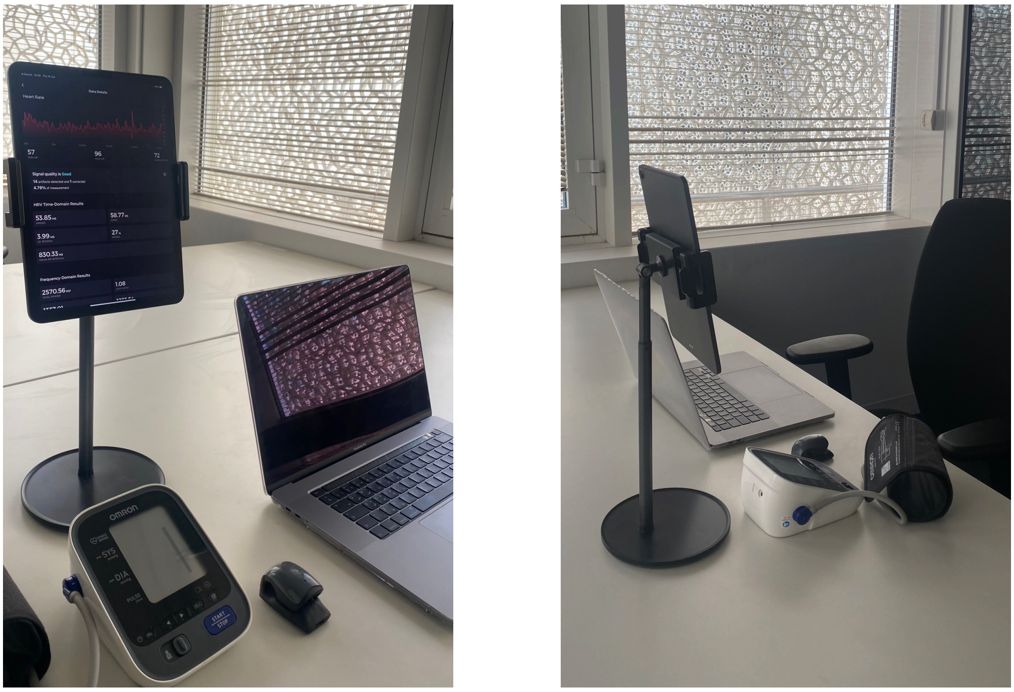
Signature of Legally Authorized Representative (When Appropriate) **Date of Signature**

Relationship to Participant (When Appropriate)

Figure E.2: Consent Form



(a) A panoramic view of the HCI Lab



(b) Experimental Setup

Figure E.3: Photos of the study room and the experimental setup at HBKU.

Table E.1: Baseline Scores and Log-Transformed HRV Measures by Group (N=38)

	Total mean (SD)	CTRL mean (SD)	HRVB mean (SD)
Affective State			
Stress	1.76 (0.97)	1.95 (1.13)	1.58 (0.77)
Mood	3.5 (0.69)	3.63 (0.5)	3.37 (0.83)
Attentiveness	15.03 (2.53)	15.37 (2.91)	14.68 (2.11)
Fatigue	7.47 (2.48)	6.37 (1.8)	8.58 (2.61)
Serenity	11.26 (2.48)	11.47 (2.61)	11.05 (2.39)
Time-Domain			
MeanRR	6.65 (0.13)	6.64 (0.11)	6.67 (0.15)
RMSSD	3.57 (0.7)	3.69 (0.71)	3.44 (0.68)
SDNN	3.97 (0.51)	3.98 (0.52)	3.96 (0.51)
pNN50	2.35 (1.31)	2.49 (1.4)	2.22 (1.24)
Frequency-Domain			
LF Power	6.47 (1.21)	6.65 (1.33)	6.3 (1.1)
HF Power	6.24 (1.39)	6.27 (1.45)	6.22 (1.37)
LF/HF	0.23 (0.95)	0.38 (0.87)	0.08 (1.03)
Total Power	7.62 (1.04)	7.73 (1.11)	7.52 (0.98)
Non-Linear Methods			
SD1	3.5 (0.71)	3.53 (0.68)	3.47 (0.75)
SD2	4.18 (0.46)	4.18 (0.48)	4.17 (0.45)
Blood Pressure			
Systolic	102.8 (13.8)	103.1 (13.7)	102.4 (14.2)
Diastolic	76.3 (6.0)	77.5 (5.6)	75.0 (6.3)

Table E.2: Pre-Intervention Scores and Log-Transformed HRV Measures by Group (N=38)

	Total	CTRL	HRVB
	mean (SD)	mean (SD)	mean (SD)
Affective State			
Stress	2.26 (1)	2.42 (1.07)	2.11 (0.94)
Mood	3.68 (0.81)	3.79 (0.63)	3.58 (0.96)
Attentiveness	16.11 (3.16)	16.16 (3.2)	16.05 (3.21)
Fatigue	6.68 (3.06)	5.79 (2.18)	7.58 (3.58)
Serenity	9.97 (2.37)	9.95 (2.32)	10 (2.47)
Time-Domain			
MeanRR	6.59 (0.13)	6.58 (0.1)	6.6 (0.15)
RMSSD	3.77 (0.58)	3.8 (0.38)	3.73 (0.74)
SDNN	3.93 (0.55)	3.86 (0.42)	3.99 (0.66)
pNN50	2.95 (0.91)	2.95 (0.79)	2.95 (1.04)
Frequency-Domain			
LF power	6.42 (1.12)	6.24 (0.85)	6.59 (1.34)
HF power	6.21 (1.53)	6.15 (1.41)	6.28 (1.67)
LF/HF	0.2 (0.81)	0.1 (0.87)	0.31 (0.75)
Total power	7.55 (1.14)	7.45 (0.94)	7.64 (1.32)
Non-Linear Methods			
SD1	3.59 (0.61)	3.48 (0.42)	3.69 (0.76)
SD2	4.1 (0.51)	4.03 (0.37)	4.17 (0.63)
Blood Pressure			
Systolic	75.7 (7.38)	119.6 (17.6)	115.5 (13.7)
Diastolic	117.6 (15.7)	78.4 (8.1)	72.9 (5.6)

Table E.3: Mid-Intervention Scores and Log-Transformed HRV Measures by Group (N=38)

	Total mean (SD)	CTRL mean (SD)	HRVB mean (SD)
Affective State			
Stress	1.82 (0.98)	2.16 (1.07)	1.47 (0.77)
Mood	3.84 (0.86)	4.05 (0.62)	3.63 (1.01)
Attentiveness	15.39 (3.09)	14.84 (3.2)	15.95 (2.95)
Fatigue	8.18 (4.03)	6.84 (3.13)	9.53 (4.45)
Serenity	10.87 (2.97)	9.16 (2.48)	12.58 (2.41)
Time-Domain			
MeanRR	6.63 (0.12)	6.6 (0.1)	6.66 (0.13)
RMSSD	3.73 (0.66)	3.6 (0.64)	3.86 (0.66)
SDNN	4.23 (0.55)	3.89 (0.39)	4.56 (0.49)
pNN50	2.73 (1.17)	2.58 (1.29)	2.88 (1.06)
Frequency-Domain			
LF power	7.2 (1.35)	6.5 (0.82)	7.89 (1.45)
HF power	6.54 (1.32)	6.54 (1.09)	6.54 (1.55)
LF/HF	0.69 (1.1)	0.03 (0.87)	1.34 (0.92)
Total power	8.01 (1.14)	7.6 (0.85)	8.41 (1.27)
Non-Linear Methods			
SD1	3.73 (0.65)	3.62 (0.58)	3.85 (0.71)
SD2	4.46 (0.51)	4.25 (0.36)	4.67 (0.57)

Table E.4: Post-Intervention Scores and Log-Transformed HRV Measures by Group (N=38)

	Total	CTRL	HRVB
	mean (SD)	mean (SD)	mean (SD)
Affective State			
Stress	2.03 (0.85)	2.21 (0.98)	1.84 (0.69)
Mood	3.97 (0.75)	4.11 (0.66)	3.84 (0.83)
Attentiveness	16.03 (3.11)	15.37 (3.68)	16.68 (2.33)
Fatigue	7.18 (3.43)	6.37 (2.83)	8 (3.84)
Mood	3.97 (0.75)	4.11 (0.66)	3.84 (0.83)
Serenity	10.13 (2.68)	8.63 (2.36)	11.63 (2.11)
Stress	2.03 (0.85)	2.21 (0.98)	1.84 (0.69)
Time-Domain			
MeanRR	6.62 (0.12)	6.59 (0.11)	6.65 (0.13)
RMSSD	3.84 (0.74)	3.79 (0.63)	3.89 (0.85)
SDNN	3.97 (0.68)	3.88 (0.54)	4.07 (0.8)
pNN50	2.99 (1.03)	3.02 (0.92)	2.96 (1.15)
Frequency-Domain			
LF power	6.75 (1.4)	6.62 (1.13)	6.87 (1.64)
HF power	6.4 (1.75)	6.19 (1.5)	6.62 (1.99)
LF/HF	0.34 (0.8)	0.44 (0.91)	0.25 (0.69)
Total power	7.73 (1.44)	7.58 (1.08)	7.87 (1.75)
Non-Linear Methods			
SD1	3.72 (0.77)	3.64 (0.64)	3.81 (0.88)
SD2	4.17 (0.64)	4.11 (0.49)	4.23 (0.77)
Blood Pressure			
Systolic	108.1 (16.2)	117.2 (14.8)	98.9 (11.9)
Diastolic	76.3 (6.76)	78.2 (6.6)	74.4 (6.5)

Appendix F

CH7 | Supplementary Materials

Performance Metrics

Confusion Matrix

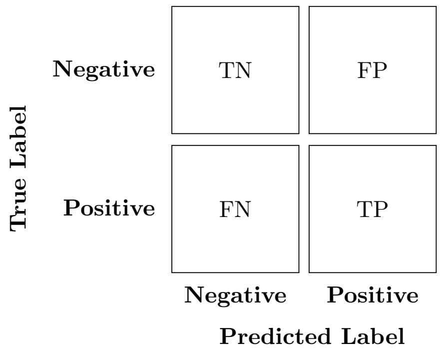


Figure F.1: Confusion Matrix
Note. TN: True Negative, FN: False Neagative, FP: False Positive, TP: True Positive

Equations

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}}$$