NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

MARIA VERÍSSIMO DUARTE JUSTINO

Bachelor of Science in Biomedical Engineering

MACHINE LEARNING MODELS FOR MENTAL STRESS CLASSIFICATION BASED ON MULTIMODAL BIOSIGNAL INPUT

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon

⟨September⟩, ⟨2022⟩

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

MACHINE LEARNING MODELS FOR MENTAL STRESS CLASSIFICATION BASED ON MULTIMODAL BIOSIGNAL INPUT

MARIA VERÍSSIMO DUARTE JUSTINO

Bachelor of Science in Biomedical Engineering

Adviser: Prof. Hugo Filipe Silveira Gamboa
Associate Professor, NOVA University Lisbon

Co-adviser: Prof. Cátia Cepeda de Vela Bastos
Associate Professor, NOVA University Lisbon

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
⟨September⟩, ⟨2022⟩

Machine Learning Models for Mental Stress Classification based on Multimodal Biosignal Input

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v6.9.5) [1].

To all the people who have loved and believed in me throughout this journey.

Acknowledgements

I used to be one of those people who wasn't sure what career path I wanted to take. My interests were broad, but I had a special taste for physics, science, and the human body; as well as a natural ability with computers, which I inherited from my two computer scientist parents. Later, when I was given entire freedom to select the subject of a high school project, I chose artificial intelligence, and ever since, I've become fascinated with this field. I decided to pursue a degree in biomedical engineering as it felt like it would provide me the opportunity to explore all these interests and aptitudes. And it did. I am proud to see that my thesis embodies my passion for science, my respect for physics, and my enthusiasm for artificial intelligence. It has been a long 5-year journey, but the people with whom I shared it made it meaningful and memorable.

First and foremost, I would like to express my gratitude to Prof. Hugo Gamboa, who believed in me and gave me the opportunity to develop my thesis at the Laboratory for Instrumentation, Biomedical Engineering, and Radiation Physics (LIBPhys). His authenticity, earnestness, and encouragement helped me grow as a student and future professional.

Additionally, I want to thank the entire LIBPhys team for their support, assistance, and companionship as I developed my thesis. In particular, I am thankful for my co-advisor, Prof.ª Cátia Cepeda, who guided me, handled my weekly progress and even with various occupations always found the time to answer my questions and discuss my work when needed. Moreover, I am also very grateful to M.Sc. Phillip Probst who was tireless throughout the development of all my thesis either explaining concepts and coming up with solutions that suited my work, or assisting me in understanding the most challenging concepts it brought.

I'm also grateful to my entire family for constantly encouraging me to set high goals and honor my abilities, even during difficult times. Special thanks to my mother, for being so caring and attentive always checking in to see how things were going and taking the time to polish with me the small details of this thesis; to my father, for always showing interest in any topic I want to explore and for actively having helped me choose the subject for this thesis; and to my grandmother, who never let me lack anything so I could

complete the course with success and health.

Furthermore, I would like to thank all the wonderful friends' life has given me. To my beloved boyfriend, Miguel, thank you for helping me throughout the coding parts of this thesis with your patience and knowledge and for supporting me during my periods of low motivation. To my remarkable best friend, Ricardo, who read once and again (and again) this document to ensure my English was flawless and who was my partner in crime during the entire course in sharing life experiences, the ones with tears and the ones with joy, thank you. Thanks also to Manena, Joana and Bernardo, three of my greatest and dearest friends who inspire me every day to strive for greater things and who never give up on their goals despite the odds.

Finally, I want to express my gratitude to everyone who is a part of my daily life but was not specifically named above. Without your love and support, I would never be where I am today. I'm hoping that even through the upcoming challenges, you'll always be a part of my life.

"A ship is always safe at shore but that's not what its build for." (Einstein)

Abstract

Mental stress is a largely prevalent condition directly or indirectly responsible for almost half of all work-related diseases. Work-Related Stress is the second most impactful occupational health problem in Europe, behind musculoskeletal diseases. When mental health is adequately handled, a worker's well-being, performance, and productivity can be considerably improved.

This thesis presents machine learning models to classify mental stress experienced by computer users using physiological signals including heart rate, acquired using a smartwatch; respiration, derived from a smartphone's acc placed on the chest; and trapezius electromyography, using proprietary electromyography sensors. Two interactive protocols were implemented to collect data from 12 individuals. Time and frequency domain features were extracted from the heart rate and electromyography signals, and statistical and temporal features were extracted from the derived respiration signal.

Three algorithms: Support Vector Machine, Random Forest, and K-Nearest-Neighbor were employed for mental stress classification. Different input modalities were tested for the machine learning models: one for each physiological signal and a multimodal one, combining all of them. Random Forest obtained the best mean accuracy (98.5%) for the respiration model whereas K-Nearest-Neighbor attained higher mean accuracies for the heart rate (89.0%) left, right and total electromyography (98.9%, 99.2%, and 99.3%) models. KNN algorithm was also able to achieve 100% mean accuracy for the multimodal model. A possible future approach would be to validate these models in real-time.

Keywords: Stress Detection, Biosignals, Occupational Health, Machine Learning, Multimodal Input.

Resumo

O stress mental é uma condição amplamente prevalente direta ou indiretamente responsável por quase metade de todas doenças relacionadas com trabalho. O stress experienciado no trabalho é o segundo problema de saúde ocupacional com maior impacto na Europa, depois das doenças músculo-esqueléticas. Quando a saúde mental é adequadamente cuidada, o bem-estar, o desempenho e a produtividade de um trabalhador podem ser consideravelmente melhorados.

Esta tese apresenta modelos de aprendizagem automática que classificam o stress mental experienciado por utilizadores de computadores recorrendo a sinais fisiológicos, incluindo a frequência cardíaca, adquirida pelo sensor de fotopletismografia de um smartwatch; a respiração, derivada de um acelerómetro incorporado no smartphone posicionado no peito; e electromiografia de cada um dos músculos trapézios, utilizando sensores electromiográficos proprietários. Foram implementados dois protocolos interactivos para recolha de dados de 12 indivíduos. Características do domínio temporal e de frequência foram extraídas dos sinais de frequência cardíaca e electromiografia, e características estatísticas e temporais foram extraídas do sinal respiratório.

Três algoritmos entitulados K-Nearest-Neighbor, Random Forest, e Support Vector Machine foram utilizados para a classificação do stress mental. Foram testadas diferentes modalidades de dados para os modelos de aprendizagem automática: uma para cada sinal fisiológico e uma multimodal, combinando os três. O Random Forest obteve a melhor precisão média (98,5%) para o modelo de respiração enquanto que o K-Nearest-Neighbor atingiu uma maior precisão média nos modelos de frequência cardíaca (89,0%) e electromiografia esquerda, direita e total (98,9%, 99,2%, e 99,3%). O algoritmo KNN conseguiu ainda atingir uma precisão média de 100% para o modelo multimodal. Uma possível abordagem futura seria efetuar uma validação destes modelos em tempo real.

Palavras-chave: Detecção de stress, Biossinais, Saúde Ocupacional, Aprendizagem Automática, Multimodal *Input*.

Contents

List of Figures

List of Tables

Acronyms

ACC Accelerometer 3, 9, 21, 27–31, 33, 34, 42, 43, 49, 53, 59, 62, 63, 75
ANS Autonomic Nervous System 18

BP Blood Pressure 18, 25
BPM Beats Per Minute 6, 46
BVP Blood Volume Pulse 18, 21–23, 25, 73

CM Confusion Matrix 13, 63, 66, 68

DT Decision Trees 11, 15, 20, 23

ECG Electrocardiogram 5, 6, 20–23, 25, 73
EDA Electrodermal Activity 21–23, 73
EEG Electroencephalogram 18, 22, 73
EMG Electromyography xxiii, 3, 5, 9, 10, 18, 22, 23, 25, 26, 33, 34, 36, 49–55, 67–69, 72, 73, 75

FLSL Fisher's Least Square Linear 22, 73
FN False Negative 13, 19, 63, 66–68, 70
FP False Positive 13, 19, 63, 64, 66–68, 70

GEE Generalized Estimating Equations 22, 73
GSR Galvanic Skin Response 18, 21–23, 25, 73

HCI Human-Computer Interaction 2, 3, 5, 11, 21, 24, 35, 36, 75, 76
HR Heart Rate xxiii, 3, 5–8, 18–23, 25, 26, 33, 34, 38, 45, 47–49, 53–55, 57, 65–67, 72, 73, 75
HRV Heart Rate Variability 5, 7, 21

KNN K-Nearest Neighbor 5, 11, 13, 16, 20–23, 26, 55, 57, 64–73, 75

LBN Linear Bayes Normal 22, 73
LDA Linear Discriminant Analysis 21
LR Logistic Regression 20, 23, 73

ML Machine Learning 2, 3, 5, 11, 15, 16, 23, 27, 54, 57, 59, 63, 75
MPL Multilayer Perceptron 22, 73

NB Naive Bayes 11, 23

PD Pupil Diameter 18, 21, 23, 73
PNS Parasympathetic Nervous System 18
PPG Photoplethysmography 6, 7, 21, 22, 34, 73, 75
PSS Perceived Stress Scale xxiii, 18, 33, 55, 64

QBN Quadratic Bayes Normal 22, 73

RF Random Forest 5, 11, 13, 15, 21, 23, 24, 55, 57, 64–73, 75
RFE Recursive Feature Elimination 55
RIP Inductive Respiration 23, 28–31, 49, 59, 62, 63
RR Respiration Rate 6, 8, 9, 22, 23, 25, 27, 29, 31, 32, 49, 55, 59, 61–63, 75
RVC Reference Voluntary Contraction 36, 44, 51

SC Skin Conductance 22, 73
SNS Sympathetic Nervous System 18, 19
ST Skin Temperature 18, 21–23, 73
SVM Support Vector Machine 5, 11, 13, 14, 20–23, 25, 26, 54, 55, 57, 64–73, 75

TN True Negative 13, 19, 63
TP True Positive 13, 19, 63

WRS Work-Related Stress 1, 2, 75

1 Introduction

1.1 Context and Motivation

Stress can be defined as an individual's physiological response to surrounding stimuli such as environmental conditions or physical exertion, resembling a feeling of strain and pressure [2–4]. It can be split into two categories: positive stress - eustress - and negative stress - distress. Eustress can be motivational and helpful performance and productivity wise. However, if too much is required from an individual's physical or mental capacity, a person's efficiency is compromised and the negative stress kicks in [5, 6]. Constant exposure to distress can deeply impact a human's physical and emotional health leading to symptoms such as headaches, cardiovascular disorders, asthma, stomachaches, diabetes, irritable bowel disease, sleep deprivation, burnout, and cancer [2, 3, 7]. Other conditions may arise from a psychological standpoint, such as depression and anxiety. This will eventually lead to difficulties regarding personal, professional, family, social, and economic affairs [8, 9].

Technology is completely interwoven into our daily activities. People spend substantial time using their phones, laptops, and other machines, considering they are widely utilized in professional and recreational activities [5, 10]. Computer careers may require significant cognitive processing and mental focusing and tend to be more desk-bound with little physical activity demand [11].

As a result of long sedentary work routines that involve extensive interaction with computer systems Work-Related Stress (WRS) disorders are becoming more prevalent among these working populations. WRS is the second most severe health issue related to work in Europe, after musculoskeletal diseases. Mental and physical stress can contribute to the onset of WRS. In 2013, the EU spent around 25€ billion as a consequence of WRS [6, 12]. Recent studies revealed that 51% of European workers experience stress in their workplaces and it is estimated that 50–60% of work absences in the business sector were caused by WRS [6, 12]. The American Institute of Stress announced that the expenses on stress-related disorders were around 300$ billion per year and, in 2015, 77% of people frequently experienced physical stress symptoms which negatively impacted on variety

life aspects [6, 12]. Twenty seven studies that included over 600 000 individuals across Europe, Japan, and the USA came to the conclusion that WRS is connected with a fairly elevated risk of coronary heart disease and stroke [13]. In comparison with people who work in a stress-free environment, the excess risk for exposed individuals is 10–40% [14].

The onset of WRS within workers can have clear repercussions on their individual work performance and well-being. The workload and requested labor hours can be high [12], making it increasingly difficult to accomplish a decent work-life balance [11]. Figure 1.1 illustrates that working conditions (and environmental circumstances) can lead to disorders. If load exceeds the person's capacity and there's a misfit between all the demands for a prolonged time period, this strain may produce serious musculoskeletal disorders [15]. With that in mind, Human-Computer Interaction (HCI) could be improved to manifest empathetic traits and adjust to its user's needs. Therefore, enabling technology to be as relaxing and pleasing as possible is a way to decrease the risk of developing a work-related disease [5].

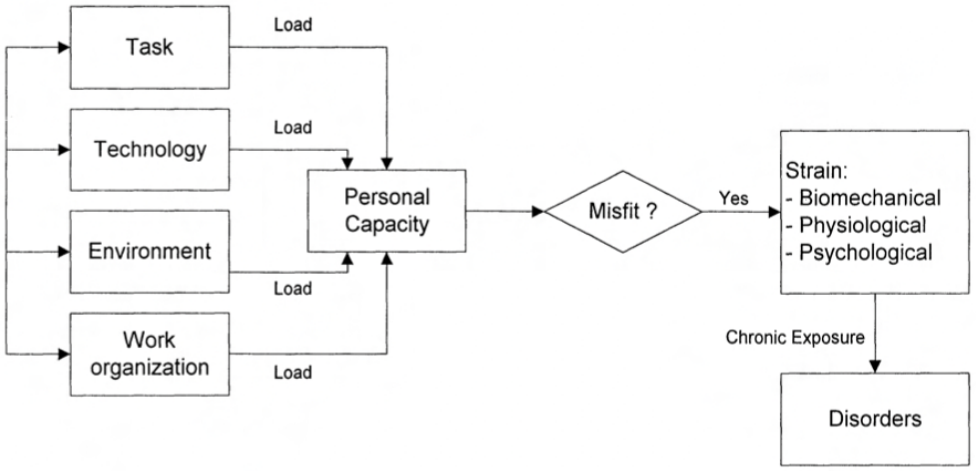


Figure 1.1: Balance Model of Work System Misfit. Retrieved from [15].

The stress level detection of a computer user can improve the computer's capability to respond intelligently, which in turn will allow the user to mitigate negative emotional states [10, 14]. Thus, approaches that monitor and access human mental states in a non-intrusive and non-invasive way can be beneficial.

1.2 Aim and Thesis Purpose

The development of this work relied on the detection of a computer user's mental state identified as 'stressed' or 'not stressed' using sensor-based measurements of physiological signals and Machine Learning (ML) algorithms. In this thesis a data collection tool entitled *Latent* [16] was used to monitor and record user interactions on a browser page in order to track their performance during protocols they completed.

Sensor data such as HR, respiration (tracked with an Accelerometer (ACC)), and EMG signal were extracted using devices such as smartphones, smartwatches, and proprietary EMG sensors. These biosignals were pre-processed in order to extract significant features and the best ones were then selected and used for the classification of numerous models, including a multimodal one. This workflow is represented with a block diagram in Figure 1.2.

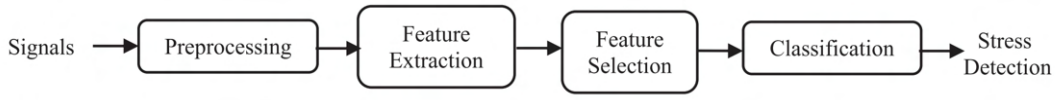


Figure 1.2: Block diagram of the proposed stress detection method. Adapted from [17].

The presented work focused on using either smartwatches, wearable sensors, or a combination of these. While smartwatches are widely used devices, most other wearables are highly specialized equipment associated with higher costs. This work explored an acquisition system that relied mostly on smart devices utilized by everyday workers: smartwatches and smartphones. Unimodal models for each biosignal were developed as well as a multimodal approach, to determine the classification capabilities between these.

Computer mouse-tracking has been proving itself to be an easily approachable method of accessing underlying cognitive processes with great richness and depth [18]. A secondary goal of this thesis was to verify if there were any distinct behaviors in the computer user's interaction under stress using the HCI data provided by *Latent* [16]. As previously stated, computers have reached widespread usage [6]. Therefore, passive real-time monitoring of stress using HCI could provide instant feedback and allow for an early intervention [9].

This thesis was developed in the Physics Department of the Nova School of Science and Technology | FCT - NOVA, together with the Biosignals team within the PrevOccupAI project of the LIBPhys group.

1.3 Thesis Structure

The six Chapters that make up this thesis are arranged as follows: the current Chapter provides the context and motivation for this thesis's subject along with an explanation of its primary goals. The theoretical background important to comprehending this thesis, including HR, respiration, EMG, HCI, ML, and stress's multimodal nature is covered in Chapter 2'. A survey of the literature on physiological, behavioral, and combined stress classification is done in Chapter 3. In Chapter 4, information regarding equipment and the techniques utilized during the thesis' development in terms of data cleaning and processing, as well as the implemented protocols, and created ML models, is presented. The work's results are reported and discussed in Chapter 5. Finally, in Chapter 6, the overall conclusions and potential future studies are addressed.

2

Theoretical Concepts

This chapter includes all key ideas needed to better understand the remaining chapters of this thesis. Starting with the physiological signals that were employed to serve as base for the thesis: HR, respiration, and EMG signals. Subsequently, the HCI concept is defined and a brief introduction to ML, with a more detailed explanation of the three algorithms that were used: Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbor (KNN) is presented. Finally, a conceptual explanation of stress's multimodal nature is provided.

2.1 Heart Rate Signal

An Electrocardiogram (ECG) is a graphic representation of the hearts electrical activity during the cardiac cycle [19, 20]. The five waves that compose an ECG are P, Q, R, S, and T as shown in Figure 2.1. Right and left atria contraction, or atrial depolarization, is what causes the P wave to be formed. It typically lasts 80–100 ms and has an amplitude of 0.25 mV. The formation of the QRS complex is brought on by the contraction of the right and left ventricles, also referred to as ventricular depolarization. Ventricular rate can be determined by the time elapsed between QRS complexes. R wave has an amplitude of almost 1.6 mV and the Q wave is approximately a quarter of that amplitude. The P-R interval lasts between 120 and 200 milliseconds. Finally, the T wave is registered when the ventricles recharge, a process known as ventricular repolarization. T waves typically last 160 ms and have an amplitude of 0.1 to 0.5 mV. Occasionally the T wave is followed by a tiny positive U wave. Repolarization of the interventricular septum or the final traces of ventricular repolarization are responsible for this wave [19–21].

Since the QRS complex is the most obviously discernible waveform from the ECG, the heartbeat is registered by detecting the QRS complexes. Thus, the R peak is used to extract information such as RR intervals and Heart Rate Variability (HRV), which are utilized to define cardiac activity. The RR interval is defined as the period between two successive R peaks and the HRV measures the oscillation in time periods between consecutive heart cycles. These can be used to calculate time, frequency, and non-linear features.

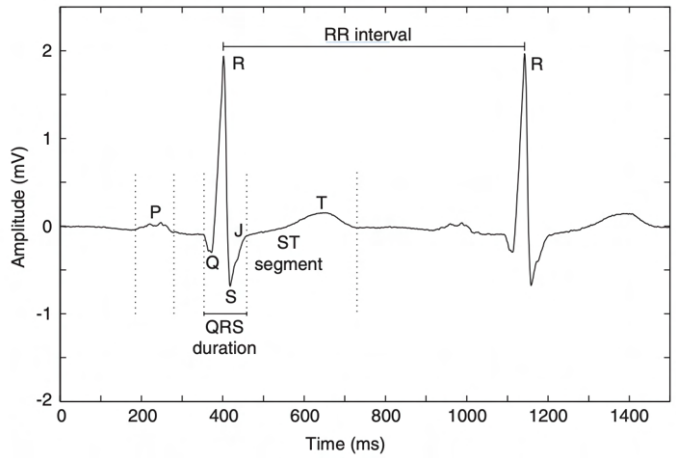


Figure 2.1: Wave definitions of the cardiac cycle. Retrieved from [22].

RR Intervals and Beats Per Minute (BPM)'s mean, standard deviation, and root mean square are a few examples of temporal domain features. The frequency domain features include very low-frequency, low-frequency, and high-frequency components. The most prevalent non-linear features include entropy, complexity, poincare plots, recurrence, and fluctuation slopes [6].

An adult's HR is typically 72 BPM. The intervals of heart experienced by different age groups are displayed in Table 2.1. Also a person's fitness is also directly linked to their HR. An athlete can have a HR as low as 40 BPM whereas a healthy person will have a resting HR of 50-60 BPM [19, 23].

Table 2.1: Heart Rate in age groups.

	Heart Rate (beats per minute)
Infants	> 100
Adolescents	60-100
Elderly	60-80

Electrocardiographic signals can provide crucial indicators of a person's health status. As a result, they are frequently acquired for a wide range of medical applications, such as cardiovascular abnormalities diagnosis, arrhythmias detection, physiological responses, sleep apnea diagnosis, chronic patient monitoring, sudden cardiac arrest forecasting, and biometric, emotional, and physical activity recognition systems [21, 24].

An ECG is the gold standard method to measure HR. However it involves electrode wiring, which is unsuitable for long-term use [25]. In comparison to the ECG, the Photoplethysmography (PPG) sensor incorporated in the smartwatch provides a less obtrusive and more practical alternative method for estimating HR or other parameters such as arterial oxygen saturation levels and Respiration Rate (RR) [26]. The sensor emits a light

and body tissues can reflect and partially absorb it. The PPG's sensing system also detects the scattered light that emerges through the skin. PPG can thus measure changes in blood flow. Compared to other tissues, blood has a higher absorption coefficient [27]. The amount of light absorbed is proportional to the amount of blood present, and the blood volume in the artery varies with the cardiac cycle. The instantaneous blood volume level can be used to determine the exact point in the cardiac cycle that the PPG sensor is measuring [25, 28]. Thus, beat-to-beat variations, known as HRV can be derived from raw PPG data [26]. Since this signal is derived from the pulse, the appropriate terminology for it would be pulse rate or optical heart-rate [26, 29]. However, for simplification purposes, the pulse rate will now be referred to as HR. Figure 2.2 shows a graphic illustration of the HR recorded with the PPG smartwatch sensor, where each dot is an estimate of the HR provided by the sensor. The smartwatch recorded HR for a total of five times in one acquisition with approximately two-minute intervals between recordings.

Figure 2.2: Photoplethysmography heart rate acquisition over time.

Due to its pervasiveness and ease of wearable integration, PPG is becoming more popular in wrist-worn devices, such as smartwatches and wrist-based fitness bands, to be employed in monitoring systems for healthcare and biomedical purposes. Its technological and practical advantages allow an easier pulse rate measurement [26, 29]. However, one of the main issues with PPG-based monitoring approaches is the high susceptibility to motion artifacts brought on by daily movements, therefore the signal needs to be processed to effectively remove this noise. Additionally, a variety of other elements, such as ambient optical noise, may influence the PPG signal capture [30].

2.2 Respiratory Signal

Respiratory measurements are frequently overlooked regarding their significant impact on HR [25]. One of the most crucial vital signs that reflects the health of the body is the RR, also known as breathing rate. It is the amount of breaths (or movements) per minute indicating the inspiration and expiration of air. Breathing patterns by age groups are displayed in Table 2.2.

Table 2.2: Respiration rate in age groups.

	Respiration Rate (breaths per minute)
Infants	30-60
Adults	12-18
Elderly	12-28

A few disorders that can alter a person's normal RR are asthma, lung illness, and congestive heart failure. Mental states, such as anxiety, can also disturb a regular RR [31].

Figure 2.3 shows a normal respiration signal. The signal amplitude grows until it reaches its peak at the end of inspiration and drops while expiring, reaching its minimum when the expiration is finished. The respiration signal is highly affected by the current emotional state of a person as shown in the arbitrary emotional respiration signal displayed in Figure 2.4. Therefore, respiration signals can be used for diagnosis: during stress, breathing becomes inconsistent and RR rises [32].

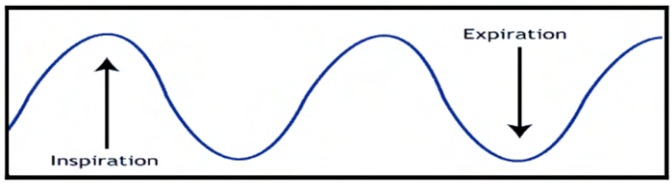


Figure 2.3: Normal Respiration Signal. Retrieved from [32].

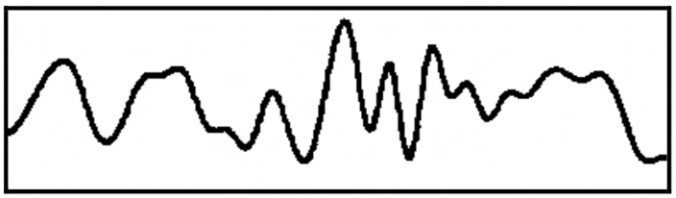


Figure 2.4: Emotional Respiration Signal. Retrieved from [32].

The patterns of breathing can be recorded using a variety of sensor devices [25]. The three primary categories of breathing-monitoring equipment are airflow quantifiers, chest and/or abdomen motion trackers and gas-level gauging devices for blood or exhaled

air [33]. Unfortunately, the majority of these technologies are either expensive or not accessible to most people [34].

As a result of the market spread of mobile technologies, numerous strategies are being investigated for the recording of vital signs, such as the RR. According to certain studies [31, 33–35], it may be possible to track respiration and calculate the RR using smartphone built-in sensors like accelerometers. Chest motion caused by the diaphragm's expansion and contraction is synchronized with lung movement, and the variation in the ACC's may accurately record it. The elimination of motion artifacts and the detection of spurious peaks are the main issues of this detection method [31]. This work tried to use a smartphone ACC to capture respiration through chest movement.

2.3 Electromyography Signal

The EMG signal, which represents neuromuscular activity, is a biomedical signal that measures electrical currents produced in muscles during their contraction. The nervous system always regulates muscle contraction and relaxation. The electromyographic signal is, thus, a complex signal that is regulated by the nervous system and depends on the anatomical and physiological characteristics of the muscle [36].

Skeletal, smooth, and cardiac are the three different types of muscle tissue that may be distinguished based on their structural characteristics, contractile ability, and regulatory mechanisms. EMG is used to analyze skeletal muscle. This muscle is attached to the bone, and it's contraction enables the skeleton support and movement [36].

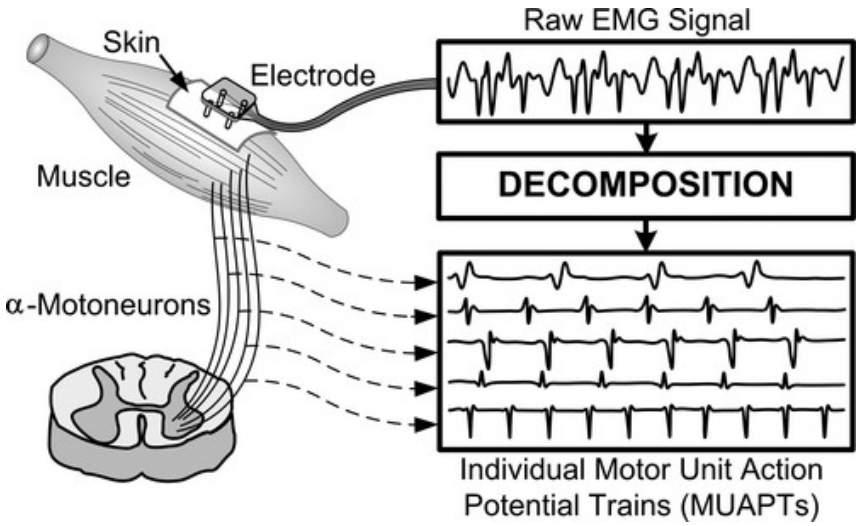


Figure 2.5: Electromyography signal and decomposition of MUAPs. Retrieved from [36].

Muscle action potential refers to the electrical signals that are conducted by muscle fibers similarly to how nerves operate. To measure the motor unit action potential, which is the sum of all the muscle action potentials from a single motor unit, a skin-surface electrode (non-invasive) placed close to the desired field or a needle electrode (invasive) can

be applied [6, 36]. The EMG signal is a combination of all muscle fiber action potentials happening in the muscles as a response to neural activity. The intervals between these action potentials are unpredictable. The EMG signal could be positive or negative at any given time with an absolute amplitude ranging between 0-10 mV ± 5mV. The process to obtain this signal and its decomposition in action potentials are depicted in 2.5 [36].

There are two key concerns that affect the quality of the signal. The first one is the signal-to-noise ratio (quotient between the electromyographic energy and the noise signal energy). This noise can come from electronic equipment, surrounding environment, motion artifacts or signal instability. The other drawback is the signal distortion. Since the EMG signal's frequencies may contain important information, caution is required when applying filters so that signal distortion can be as minimal as possible [36, 37].

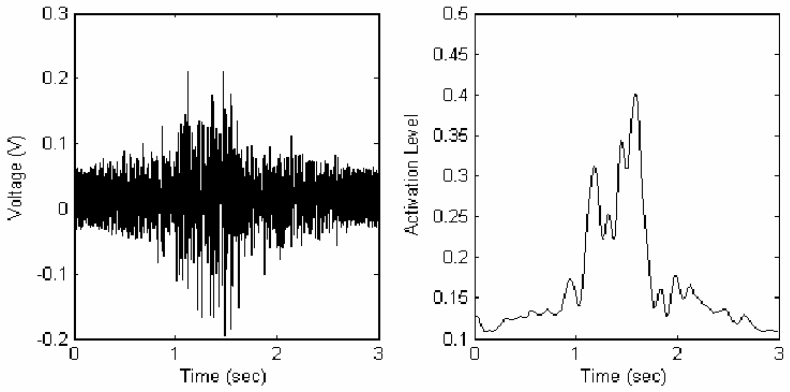


Figure 2.6: Raw electromyographic signal and corresponding activation level. Retrieved from [38].

The use of reliable and advanced methods for the EMG signal identification in clinical diagnosis and biomedical applications such as rehabilitation, prosthesis control, muscle fatigue analysis, and clinical diagnosis is growing [36, 37]. To further analyze the EMG signal, time and frequency domain properties such as peak loads and gaps, root mean square, mean absolute value, variance, energy, mean, and median frequency, zero-crossing, and frequency ratio can be retrieved. The moments of muscle activation over a period of time can also be observed and features can be extracted from it. Figure 2.6 on the right illustrates a muscle activation period that corresponds to the EMG signal on the left [6, 17, 36].

2.4 Human-Computer Interaction

There are several definitions of HCI, the most comprehensive being the study of the interaction between humans and computers [39, 40]. More specifically, HCI can be considered the field of assessment regarding how the design, evaluation and implementation of interactive computing systems, such as computers, mobile phones, household appliances, among others, influence work and human activities. It also uses knowledge from other research fields such as psychology and sociology to achieve its main goal of improving the relationship between human and computer through the increase of usability and functionality towards the user's needs [39, 41].

A user's emotional state can be recognized through HCI using audio, text, video, mouse cursor motion, etc. [42, 43]. This thesis will be using *Latent* [16], a data collection tool providing all the above options and more.

2.5 Machine Learning

ML is considered to be a branch of artificial intelligence that studies how computer algorithms can make accurate predictions (or reactions) in specific conditions. A ML model receives a sequence of data inputs and executes algorithms that learn in a supervised, unsupervised, semi-supervised, or reinforced way [44, 45].

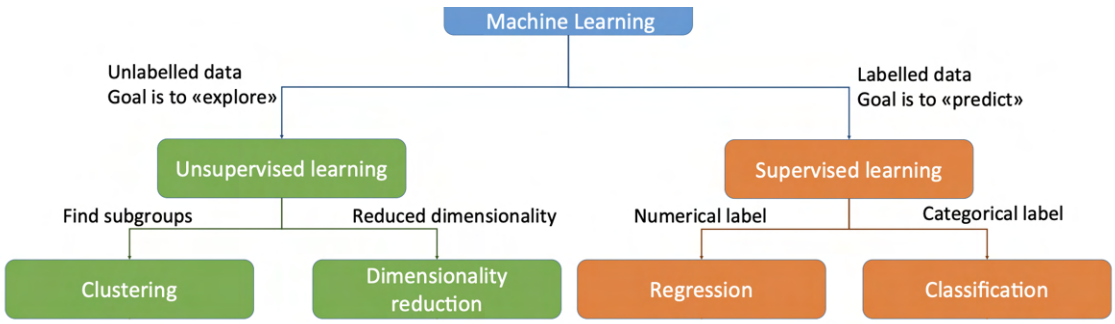


Figure 2.7: Taxonomy of the different machine learning methods presented. Adapted from [46].

The supervised and unsupervised approaches can be distinguished by their main goal. The first one aims to explore the received data while the latter wants too make predictions out of it. Figure 2.7 offers a scheme where the distinction of the two approaches can be clearly seen.

In supervised learning, illustrated on the left side of Figure 2.8, a sequence of outputs is provided in addition to the input sequence, and the goal is to learn how to achieve one of the desired outputs from a new input. This method is divided into two categories: regression (output is a real number) and classification (output is a class label)[47]. Decision Trees (DT), RF, SVM, KNN and Naive Bayes (NB) are examples of well-known supervised classification models. Some notable regression models include linear regression,

step-wise regression, ordinary least squares regression, multivariate adaptive regression spline, and locally estimated scatter-plot smoothing [48].

Unsupervised learning methods, visually represented on the right side of Figure 2.8, simply receive an unlabeled training data-set and attempt to deduce existing patterns or clusters from it. The four areas of unsupervised learning methods are clustering, dimensionality reduction, anomaly detection, and association rule mining [45, 47].

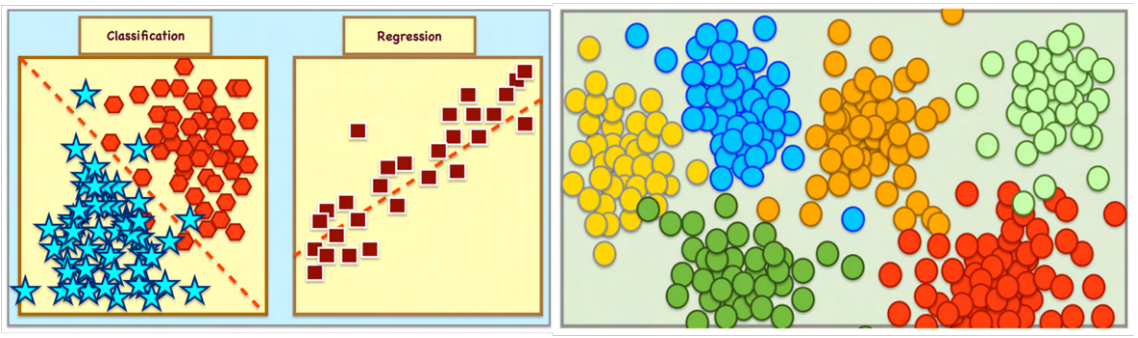


Figure 2.8: Left: Supervised learning. Right: Unsupervised learning. Retrieved from [49].

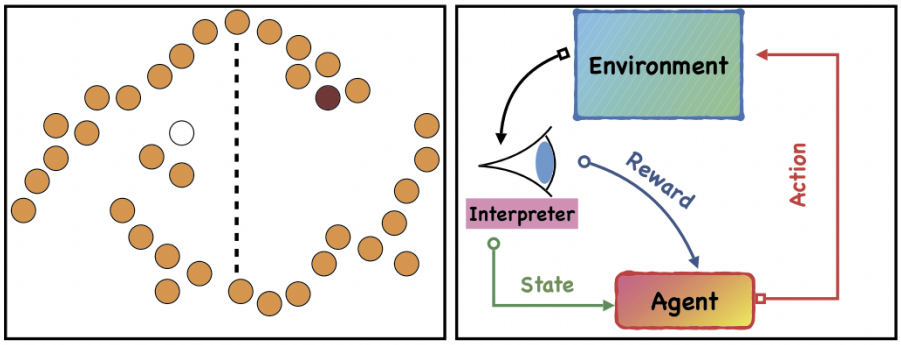


Figure 2.9: Left: Semi-supervised learning. Right: Reinforcement learning. Retrieved from [49].

Semi-supervised algorithms, presented on the left side of Figure 2.9 combine supervised and unsupervised techniques. They are used to clarify learning situations with a limited amount of labeled data and a large amount of unlabeled data. Having a mixed data-set is quite frequent, especially in domains where acquiring data is accessible but labeling it is costly or time-consuming [47].

Additionally, there are reinforcement learning algorithms (shown on the right side of figure 2.9) where the algorithm interacts with an environment by producing actions that will influence its state. As a result, the algorithm obtains some scalar rewards (or punishments), with the final aim of learning how to maximize those future rewards (or reduce punishments) over the course of its existence [47].

In order to accomplish some emotion differentiation (stress from not stressed), this

thesis used classification supervised learning models. When classifying an unknown sample, these models can perform a correct or incorrect prediction. When classifying a sample, the class where the sample belongs is considered positive and the remaining ones are negative. Consequently, these concepts serve as the foundation for the following created evaluation metrics:

- True Positive (TP): Number of correctly predicted positive sample classes.
- True Negative (TN): Number of correctly predicted negative sample classes.
- False Positive (FP): Number of incorrectly predicted positive sample classes.
- False Negative (FN): Number of incorrectly predicted negative sample classes.

Table 2.3 shows a Confusion Matrix (CM) which is commonly used to better visualize the above concepts.

Table 2.3: Confusion Matrix.

		Predicted Class	
		No	Yes
Actual Class	No	True Negative (TN)	False Positive (FP)
	Yes	False Negative (FN)	True Positive (TP)

To evaluate a trained model, there are some evaluation metrics commonly used, such as sensitivity or recall 2.1, specificity 2.2, positive predictivity or precision 2.3, negative predictivity 2.4 and accuracy 2.5 [24, 50, 51].

$$\text{Sensitivity/Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2.1) \qquad \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (2.2)$$

$$\text{Positive Predictivity/Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2.3) \qquad \text{Negative Predictivity} = \frac{\text{TN}}{\text{TN} + \text{FN}} \quad (2.4)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (2.5)$$

As will be stated in Chapter 3, many models were commonly employed in investigations using physiological signals to detect stress. According to the findings in the literature, the most typically used classifiers are SVM, KNN, and, occasionally, RF. As a result, to maintain consistency with past investigations, these same classifiers were evaluated, with the expectation of favorable findings. The following sections (Sections 2.5.1, 2.5.3. 2.5.2) provide a quick description of how these algorithms work.

2.5.1 Support Vector Machine

The fundamental idea behind a SVM algorithm is to establish lines or hyperplanes in an N-dimensional space that define the decision boundaries by dividing the labeled training data into classes [52].

Depending on the application, the following processes are taken in order to choose the best line/hyperplane:

1. Lines are drawn using kernel functions to divide data points into distinct classes;
2. Data points from both classes that are closest to a certain line or hyperplane are identified, and these points are referred to as "support vectors";
3. The margin (distance between that line/hyperplane and the support vectors) is then calculated;
4. The line/hyperplane that maximizes the margin is considered the best and therefore is chosen.

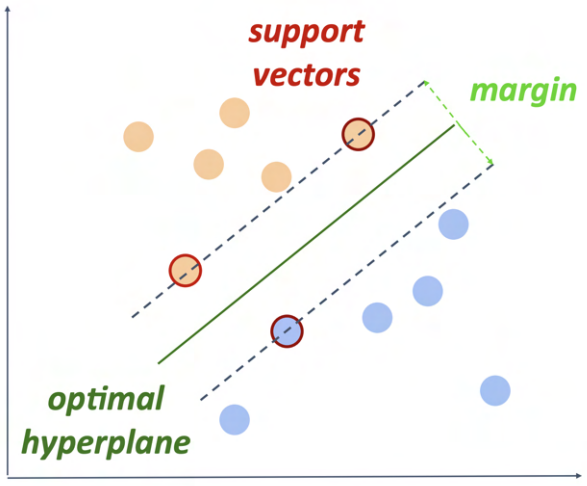


Figure 2.10: Hyperplane separating the support vectors from each class. Retrieved from [46].

Figure 2.10 helps the visualization of both the concept of "support vectors" and the margin between them and the line/hyperplane. The region of the hyperplane where the new unknown sample is better fitted decides how it should be classified. This classifier is not the only one employing lines/hyperplanes for classification. However, the SVM differs from the others due to using the maximum margin separating hyperplanes. By choosing an optimal hyperplane using this method, the SVM's capacity to correctly forecast the classification of previously unobserved instances is maximized [53].

2.5.2 Random Forest

The fundamental idea behind a RF algorithm is to forecast the outcome variable by using a set (ensemble) of trained DT. The classification of this algorithm is based on the most frequently occurring prediction among the wide collection of individual DT. This algorithm works well due to its low correlation between different DT, making ensemble predictions more accurate than each individual tree's prediction [54].

DTs in their original form are seldom ever employed in ML because they are prone to overfitting, which is why each individual DT used for the construction of the RF classifier is probably overfitted. In spite of this, by combining the outputs of the multiple trees used, one can solve the overfitting problem. Each DT is composed by two components: branches and nodes. The fundamental idea behind creating a DT is to gradually evaluate various features and select the one that best divides the training data at each node [46]. Figure 2.11 displays an example of how DT operate. The graph's initial node, known as the root node, typically assesses the feature that best divides the data; subsequent nodes, known as intermediate nodes, test additional features; and finally, the terminal nodes of the tree, known as the leaves, are the ones that make a prediction. Once the DT is built, the new sample is submitted to tests in each node, from the root to the leaf, until the prediction is made [46]. The degree of impurity determines which feature best separates the training data set at each node. The gini impurity and entropy are two examples of impurity metrics. The number of nodes and the depth of the tree are two factors that determine the complexity of the tree, which in turn affect the output's accuracy [54].

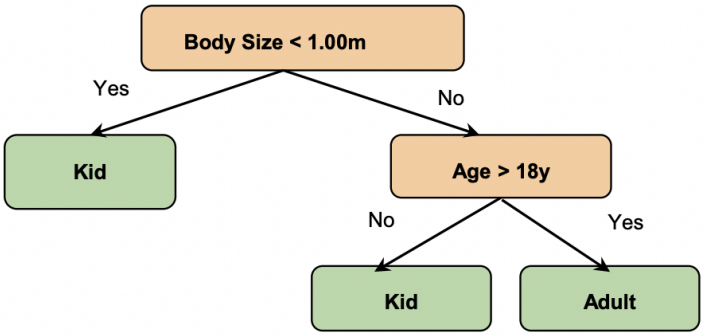


Figure 2.11: Decision Trees approach. Retrieved from [46].

2.5.3 K-Nearest Neighbors

Techniques that completely skip the learning phase and therefore do not result in an explicit model that learns from the training data can be referred to as "instance-based learning"[46]. KNN is a ML algorithm that uses the above rule and is based on a distance approach. These approaches simply store all the labeled sets of data that already exist in a database, and when a new unclassified example is observed, the algorithm will assume that observations with similar properties already exist nearby and will typically have similar outcomes [46]. In other words, the values of the data points around a particular data point define its value as shown in Figure 2.12.

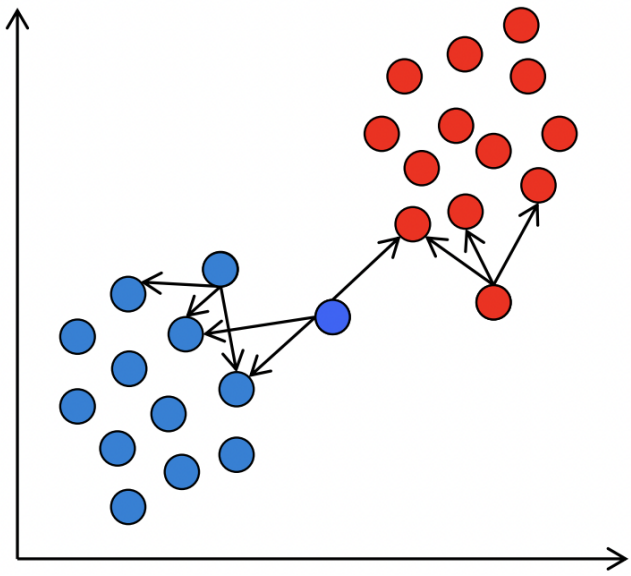


Figure 2.12: K-Nearest Neighbors approach. Retrieved from [46].

This algorithm can be described in the following six steps:

1. After loading the data, the number of neighbors to take into account (k), must be selected;

2. Then a calculation of the distance between the unknown unclassified example and all other examples from the training set is done. This distance can be calculated in many ways but in this thesis only the Euclidean (Equation 2.6 [55]) and the Manhattan (Equation 2.7 [56]) distances were used;

$$D_{\text{Euclidian}}[a,b] = \sqrt{\sum_{i-1}^{d}(a_i - b_i)^2} \tag{2.6}$$

$$D_{\text{Manhattan}}[a,b] = \sum_{i-1}^{d}|a_i - b_i| \tag{2.7}$$

3. The distance and the index of the example from which the distance was calculated are saved into a list.

4. That list is ultimately ordered so that the distances go from the smallest to the largest;

5. From this list, the first k-entries are chosen;

6. The unknown example is now classified in accordance with the those k-entries labels. In the case of regression, the label is determined by the average of the labels from the k-entries, and in the case of classification, the label is determined by the label that appears most frequently among the k-entries' labels.

The biggest drawback of this algorithm is the absence of a learning phase making it impossible to identify which features are truly important for predicting the class of a new case [46]. For instance, if 10 features were given, only 2 might be really relevant for the classification, but the distance will be computed taking all 10 into account. Thus, the k-nearest data points can be significantly impacted by irrelevant features [46].

2.6 Multimodal Nature of Stress

A clearly defined and scientifically recognized method to quantify human stress does, to this day, not exist. Stress research has focused on the study of physiological changes, but other areas, such as behavioral alterations, can also provide important information. Additionally, since context influences how people react to stress, measurable contextual information can also offer valuable hints regarding people's stress levels [57]. The multimodal nature of stress is depicted in Figure 2.13.

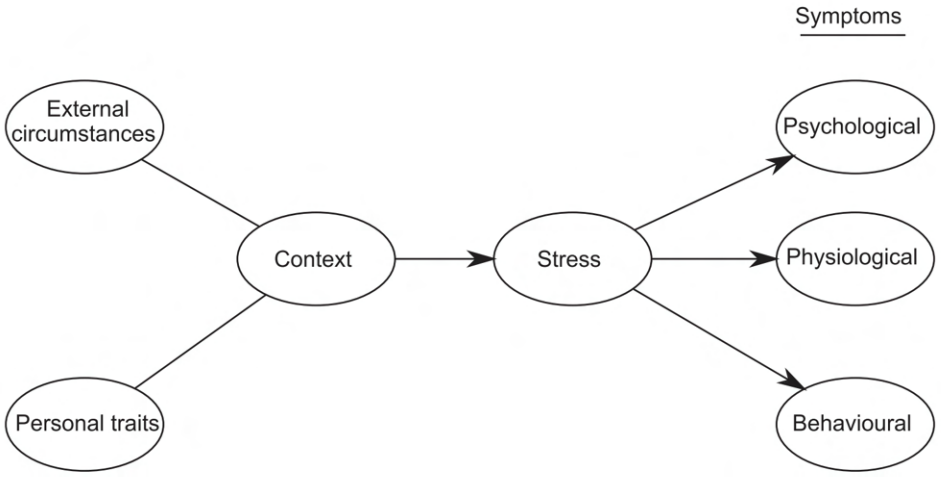


Figure 2.13: Multimodal Nature of Stress. Retrieved from [57].

Context features include personal traits and external circumstances such as schedules, calendar and ambient sound. Regarding the symptoms manifested when stress is experienced, one can categorize them into psychological, physiological and behavioral.

Psychological evaluation of stress can be carried out by means of self-report questionnaires. The PSS was utilized in this thesis. Other examples are the Stress Self Rating Scale (SSRS) [57], the Stress Response Inventory (SRI) [57], Stress Self-Rating Scale (SSRS) [6], and Positive and Negative Affect Schedule (PANAS) [6].

Physiological signals can provide information regarding the intensity and quality of an individual's internal affect experience [57]. When the body senses a challenging or hazardous situation, real or imagined, its protective mechanisms are prompted to launch a process known as "fight-or-flight response". A distress signal is then sent to the hypothalamus, which controls involuntary body tasks through the Autonomic Nervous System (ANS). The latter is divided into the Sympathetic Nervous System (SNS) and the Parasympathetic Nervous System (PNS). In the "fight or flight response", the hypothalamus activates the SNS which will release stimuli-inducing stress hormones. In opposition, the PNS branch is most active in calming and undemanding circumstances, returning the body to its resting state [3, 17]. The term "relaxation response", coined by Dr. Herbert Benson, is an opposite reaction to the "fight or flight response". Both of these responses can be observed in the individual's physiological symptoms. For example, in the "fight or flight response", HR increases, and in the "relaxation response", it decreases. In addition to HR, there is a wide range of physiological features also related to ANS activity, including Respiration, Electroencephalogram (EEG), EMG, Blood Pressure (BP) Blood Volume Pulse (BVP), Galvanic Skin Response (GSR), Pupil Diameter (PD), Skin Temperature (ST) [25, 58].

Finally, behavioral symptoms refer to expectations of how a person or a group of people will behave in a given scenario based on predetermined protocols, guidelines for conduct, or established social practices [57]. Some of the caused changes are well-known, such as feeling significantly more agitated or angry. However, these changes are difficult to quantify [57]. Other potential behavioral attitudes have been examined, for example, analyzing interactions with technology equipment can confirm its connection with stress and develop a reliable method for assessing it. Keyboard usage, text linguistics, mouse usage, speech, posture, computer exposure, smart home sensor events, facial expressions, and smartphone usage are all examples of behavioral features that can be extracted [57].

Stress responses are manifested in the three aforementioned modalities; thus, an effective stress detection system should use as much of these evidence as possible [57].

3 Related Work

The following chapter presents the related work that is relevant for this thesis. Mainly, scientific contributions on the topic of stress prediction and detection will be discussed.

3.1 Physiological Stress Classification

Early detection of physiological stress symptoms to prevent stress-related health problems has been explored. Several studies show promising outcomes on stress detection using sensor-based methods.

3.1.1 Heart Rate Stress Classification

Because the HR is so closely linked to SNS activation, it is recognized as one of the best stress markers.

Salai and colleagues [3] tested the feasibility of a low-cost chest belt HR sensor. The protocol for their study included a 10-minute relax phase followed by a provoked-mental-stress phase with the same duration. They created a simple method for detecting stress utilizing only three time-domain features of the HR signal: HR Mean (mHR), Root Mean square of successive RR differences (RMSSD), and Percentage of number of pairs of adjacent RR intervals differing by more than 50 ms to all RR intervals (pRR50). The algorithm developed by them is depicted in Figure 3.1. When the selected features progressively reached specific values, it was assumed that the subject was under stress. Their method achieved an accuracy of 74.6%, sensitivity of 75.0%, and specificity of 74.2% without the usage of machine learning. Instead, a TP outcome was considered if the technique marked the stress-inducing part as stress, a TN result if the relax part was marked as rest, an FP result if the relax part was marked as stress, and an FN result if the stress-inducing part was marked as rest. They were able to demonstrate that even a low-cost HR monitor device could allow for the extraction of features that diverge significantly under the effects of mental stress, and because they only used time-domain features, it can be implemented more efficiently (frequency-domain features require significantly more computing power to calculate).

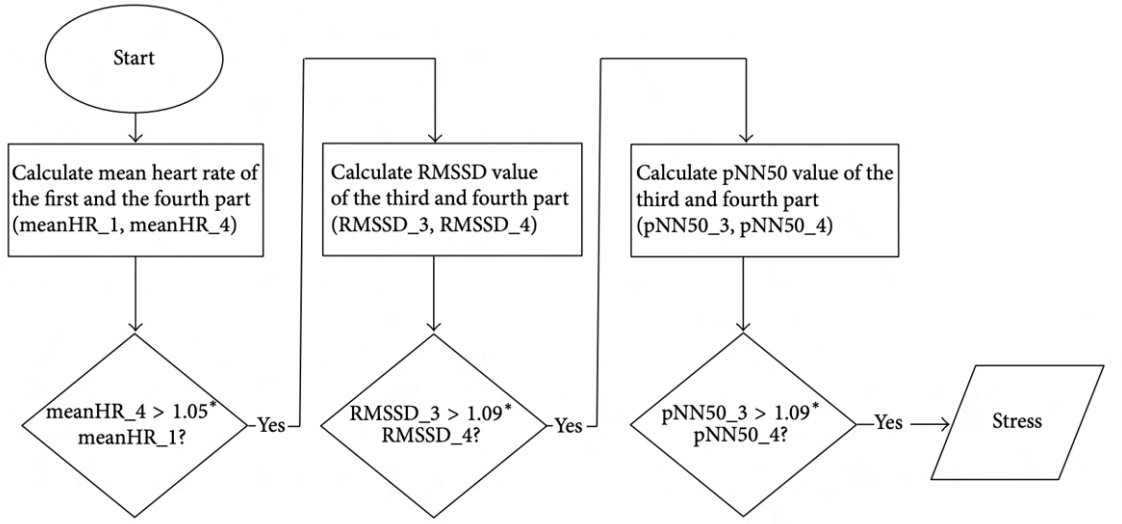


Figure 3.1: Flow chart of stress detection algorithm. A sliding window that shifted across the heart rate signal was separated into four equal portions. The algorithm detects stress if the mHR in the fourth part increases by more than 5% when compared to the first part, and RMSSD and pRR50 values decrease by more than 9% when compared to the third part. Retrieved from [3].

Karthikeyan and colleagues [59] used ECG and HR data to assess stress. The Stroop color word test was utilized as a stressor, and the signals were recorded during the experiment to discover the differences caused by it. The stress-related features were extracted from the signals using a time and frequency domain analysis. The non-linear KNN classifier yielded maximum average accuracies of approximately 94%.

Some studies opt to discreetly track a person's HR and few other features throughout the day, even during sleep, in the least intrusive possible way. Lawanont and colleagues [60] proposed utilizing an activity tracker bracelet and a smartphone to detect stress levels. The bracelet acquired several features related to HR during the day and during sleep, in addition to a few extra features about sleep cycles, calorie usage, and number of steps. They trained the model utilizing three algorithms: KNN with k = 5, SVM, and DT. Each subject's data-set was randomly divided into a training set and a test set with an 80% to 20% ratio. The highest accuracy obtained while using a DT to predict stress on the test set was 78.95%. Muaremi et al. [61], instead of using a bracelet, chose to employ a wearable chest belt in addition to a smartphone for the same purposes. They also used HR information collected during day and night and some features from audio and physical activity. They did this over the course of four months, applying a leave-one-day-out cross-validation approach for every individual separately to predict the specific accuracy of each one. With multimodal Logistic Regression (LR) techniques, they obtained an accuracy of 55% using only the extra features and 59% using just the HR features. The combination of all features resulted in a 61% classification rate for a three-stress level (low, moderate, and high perceived stress) classification scenario.

The subjects' acceptance and comfort with biomedical measuring devices are significantly influenced by how unobtrusive and non-invasive these types of equipment are [57]. Smart wearable technology is becoming more prevalent. Smartphones and smartwatches enable the collection of large volumes of data without the user's being aware of it [57]. In the healthcare context, the ECG signal has been used to measure HR. However, PPG assessment is preferable for HCI studies because it is a far less intrusive and obtrusive monitoring technique. This sensor can be easily embedded in smartwatches [10]. To identify mental stress using only the PPG signal, Kalra and Sharma [58] carried out a stress-inducing experiment interspersed with periods of relaxation. Using Deep Neural Networks they attained an overall accuracy of 91 ± 1.1%. Similarly, Zubair and associates [62] designed an experimental protocol in which PPG data was acquired. The protocol used Mental Arithmetic Tasks (MAT) to appropriately stimulate various levels of stress. Different features were extracted using HR estimated from PPG signal 60-second segments. The outcomes of the proposed approach achieved 94.33% accuracy using SVM for five-level mental stress discrimination.

3.1.2 Multimodal Stress Classification

The multimodal physiological nature of stress, in addition to the extensive research in this field, suggest that more accurate models should incorporate many modalities to assist in distinguishing between stress and non-stress scenarios [57].

The HR is not the only signal that smartwatches can provide. This unobtrusive device is used in numerous studies to gather multimodal physiological information for stress classification. Zhai et al. [10] tried using BVP, GSR and PD retrieved from a smartwatch's PPG sensor to monitor the stress felt by computer users. Using a SVM to differentiate the stress state from the normal working state based on recorded physiological signals, they achieved a maximum classification accuracy of 80%. They suggested that additional classification techniques should be investigated in order to identify the mental state of stress with even greater accuracy. Moreover, Ciabattoni et al. [2] used GSR, HR and ST acquired by a commercial smartwatch and implemented a KNN classifier with 1 neighbor. To smooth data from noise and movement artifacts, the smartwatch's obtained features were filtered and interpolated. They used 10-fold cross-validation and the KNN algorithm yielded, on average, an accuracy of 84.5% when predicting if the participant was stressed or not. Additionally, a 26% misclassification error was reported when attempting to identify stress in a relaxed person. Siirtola and associates [11] also used a smartwatch for the same purpose. They provided ACC, BVP, Electrodermal Activity (EDA), HR, HRV and ST signals. The fundamental goal of this study was to determine if it was possible to properly identify stress using sensors incorporated in commercial smartwatches, not including the EDA sensor. Three classifiers — Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and RF — were used to test various window sizes and feature combinations. The leading outcome was obtained with a Linear Discriminant Analysis

(LDA) classifier using a combination of ST, BVP, and HR. They determined that models were more accurate when trained without the EDA signal and were able to achieve an accuracy of 87.4% using the leave-one-out technique.

According to research, PPG also allows multilevel stress classification. As mentioned in the section above, Zubair et al. [62] were successful in differentiating 5 levels of stress utilizing features extracted from the HR signal provided by the PPG sensor. Majid and associates [63] combined three distinct classifiers in their work to classify human stress levels using selected features from EEG, GSR, and PPG. They attained an accuracy of 95% for two classes—non-stressed and stressed—and a 77.5% for three classes — non-stressed, mildly stressed, and highly stressed — using a Multilayer Perceptron (MPL) classifier and a combination of features from all three modalities.

Some research aimed to study the relevance EMG could have in stress detection. Karthikeyan et al. [64] used trapezius EMG in a stimulated stress-inducing environment to study the association between changes in human stress levels and muscular tension. For stress classification, a basic non-linear classifier KNN was employed and a maximum average classification accuracy of 90.70% was obtained. In 2011, Wijsman and colleagues [65] used trapezius muscles' EMG, respiration, Skin Conductance (SC), and ECG to identify mental stress. They distinguished between stress and non-stress conditions using multiple classifiers, including Linear Bayes Normal (LBN), Quadratic Bayes Normal (QBN), KNN, and Fisher's Least Square Linear (FLSL), and achieved an almost 80% consistent accuracy for all of them. In 2013, some of the same authors from the previous study [66] investigated if the trapezoids were suitable muscles for stress detection and concluded that they were (i.e., the EMG exhibited greater amplitudes and fewer gaps - periods of relaxation - during stress compared to a resting state). Later on that year, the same authors [67] used HR, respiration, GSR, and EMG of the upper trapezius muscles to distinguish between states of stress and rest in working contexts. The protocol that was implemented included stress tests that were aimed to simulate office-like circumstances. Nineteen features were extracted including HR Standard Deviation (SDHR), Low Frequency (LF), High Frequency (HF), Symphatovagal Balance Index (SVI), RR, Root Mean Square of Successive (RMSE), Mean and Median Frequencies (MF and MDF). The classification accuracy obtained utilizing Generalized Estimating Equations (GEE) was 74.5%. All studies from this author and associates implemented the arithmetic "Norinder Test" on their stress-inducing protocols. Pourmohammadi and Maleki [17] conducted research to compare the efficiency of the EMG signal with the ECG signal in detecting mental stress. According to their findings, EMG and ECG signals can accurately diagnose stress levels with 100%, 97.6%, and 96.2% accuracy for two, three, and four degrees of stress, respectively using the SVM classifier. They also demonstrated that the EMG signal outperformed the renowned ECG signal in the stress detection field.

There are still some other worth mentioning studies that sought to classify stress using a multimodal approach. In 2006, Zhai and others [68] developed an emotion recognition method. They set up an experiment for physiological sensing, then performed signal

pre-processing to extract features, and then used a ML algorithm for emotion recognition. Four signals: GSR, BVP, PD and ST were monitored and analyzed. The supervised classification of the emotional states "stressed" and "relaxed" was carried out using a SVM algorithm, which produced satisfactory levels (up to 90.10%) of differentiation. Shi et al. [69], in 2010, presented models for detecting stress from physiological measurements. They made use of a chest band that had ECG, GSR, Inductive Respiration (RIP) and ST sensors. Experimental results showed that their models could detect stress with 56% precision for the specific models, and for the general ones, a precision of 62%; using the SVM algorithm. Choi et al. [25] conceived and created a minimally invasive wearable sensor platform allowing long-term ambulatory monitoring of mental stress, in 2011. Using a LR model, their feature set from EDA, HR, EMG and respiration sensors was able to accurately predict mental stress with an accuracy of 81.0%. Palanisamy and associates [70], in 2013, also attempted to identify human stress employing multiple physiological signals including ECG, EMG, HR, GSR, and ST and using mental arithmetic task-based stress-inducing stimuli. The subjects' normal and stressed states were distinguished using the nonlinear classifiers KNN and Probabilistic Neural Network (PNN). According to the findings, the proposed HR model performed well achieving an accuracy of up to 93.75%. The results for the ECG, EMG, GSR, and ST were, respectively, 76.25%, 71.25%, 70.32%, and 75.32%. In 2015, Smets et al. [71] compared different ML techniques for stress level prediction based on physiological responses in a controlled environment. During the stress test in the lab, the EMG, GSR, ST, and respiration were recorded. Six ML algorithms were analyzed using a general and personal approach: LR, SVM, DT, RF, NB. The results demonstrate that personalized NB and generalized SVM rendered the best average classification results with 84.6% and 82.7% respectively.

This work was build upon the research mentioned above. As multimodal models were able to classify with higher accuracies than models based on single sensor input, this multimodal approach was followed. The equipment chosen to extract these signals is widely available and simple to use. One of the extracted signals - the RR - was not extracted with a specialized equipment. This signal was instead obtained from a smartphone placed on the chest. A smartwatch was used to collect the HR signal and the EMG signal was extracted from the upper left and right trapezius muscles with proprietary EMG sensors. Furthermore, instead of striving to use Deep Learning (DL), only ML algorithms were employed because they are more light-weight and need less computational capacity.

3.2 Behavioral Stress Classification

Focusing on how HCI can help analyze human behavior instead of sensor-based tools allows the measurements to be equipment-free and even more imperceptible.

Salmeron et al. [72] aimed to evaluate the effectiveness of non-intrusive, low-cost indicators obtained from user interactions with the mouse and keyboard in identifying affective states and behavior changes. They attained an accuracy of 59% using an RF classifier after extracting 96 features from mouse interaction and 42 features from keyboard usage. Sun, Paredes and Canny [73] explored stress measurement from common computer mouse operations. Increased muscular activity happens when the human body experiences stress. Therefore, they decided to use common mouse operations to directly detect muscle stiffness of hand and arm movement. They found that click force was significantly higher during a stressful moment and stated that motion control features were remarkably different in stressful situations and non-stressful ones. An approach developed by Gonçalves and others [8] tried to access the stress of computer users during demanding tasks, in an unobtrusive way, through mouse and decision-making behaviors. When the user manifested greater stress levels, probably nearly reaching the state of burnout, his/her performance diminished. This is recognized by the evidence of less efficient interaction patterns - extended mouse clicks, greater distances traveled by the mouse cursor, larger keystroke downtimes, accelerated and diminished decisions and counting.

Other studies aim to reduce stress after its detection through HCI. A review by Daher and others [5] investigated the influence of blue-colored light in negative stress reduction. Figure 3.2 shows the architecture behind their work. They stated that, in comparison to the normal state, the experiment with no blue light induced more stress than the experiment with blue light, which implied blue-colored light tends to reduce mental stress.

Saxena et al. [4] used the pressing rate on the keys of a keyboard; mouse trajectory, mouse speed rate and the time a user spends in front of a system to establish six stress levels based on how much and what HCI is occuring - "No Work", "Slow Work", "Speed Up", "Optimal Speed", ""Undergoing slow work progress", and "Abort Work". A working user usually experiences all these levels. They concluded that if a system captures which level is being experienced by the individual, it could adjust itself to the user's needs, not only increasing its usability but also reducing the user's stress level.

In this thesis, assessing the user's mental state of stress was attempted with features extracted from behavioral response, in particular those from HCI; in addition to the stress detection model created using physiological signals.

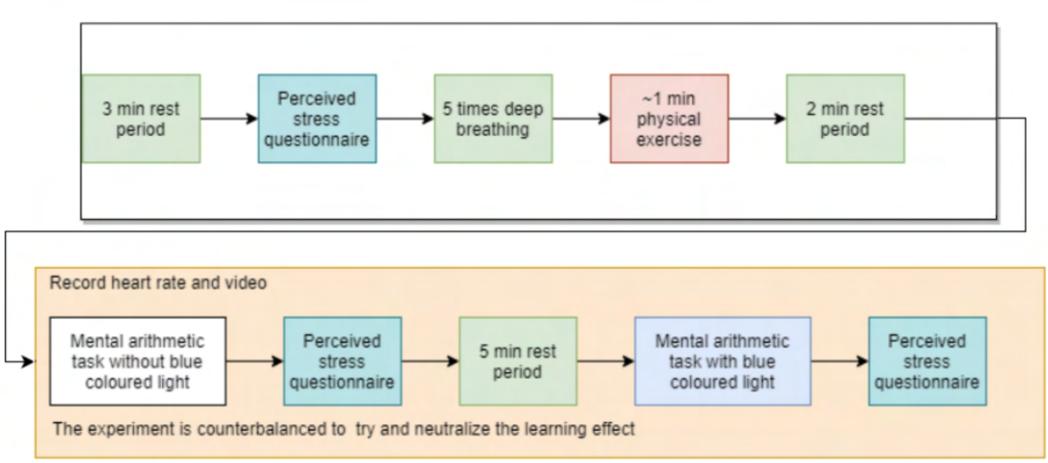


Figure 3.2: Experiment Architecture. Retrieved from [5].

3.3 Combined Stress Classification

To replicate different real-life scenarios and get more accurate findings, multiple responses to stress are preferable [74]. To successfully and instantly meet the user's needs, high detection accuracies (>90%) are required [74].

Aigrain et al. [75] obtained average F1-scores of up to 85% with the SVM classifier using multimodal features. They used physiological signals such as BVP, ECG, EMG, GSR,nHR and behavioral signals: speech, body movement and head position among others. The best results were obtained when combining both stress responses with a mean F1 score of 85.5% (± 2.0%). A more in-depth study conducted by Akmandor and colleagues [74] presented an automatic stress detection and alleviation system, called SoDA. In order to continuously monitor human stress levels and reduce stress as it develops, SoDA uses evolved wearable medical sensors, specifically the ECG, GSR, RR, BP, and Blood Oximeter. It detects stress and reduces it in a user-transparent manner. When SoDA detects stress, it applies an adaptive stress-relief approach based on the user's stress response. Figure 3.3 displays a schematic of the stress detection and alleviation process.

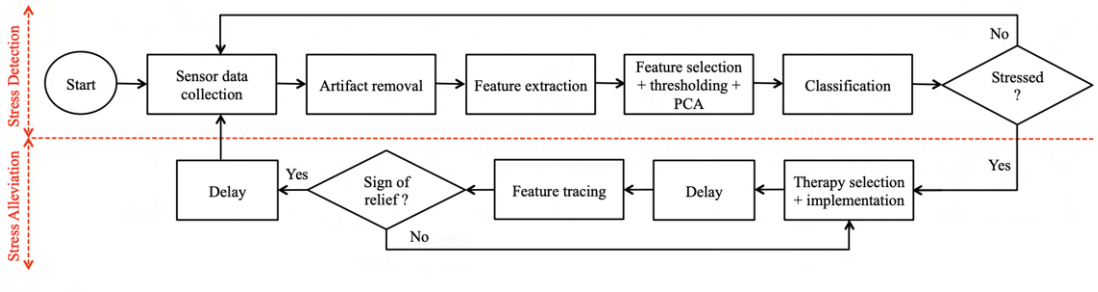


Figure 3.3: Block Diagram of SoDA system. Retrieved from [74].

They establish the effectiveness of the proposed system through a detailed analysis of data collected from 32 participants. Four stressors and three stress reduction techniques were used. With a unique fusion of supervised feature selection and unsupervised dimensionality reduction, SoDA's stress detection achieves 95.8% accuracy (KNN) and 86.7% accuracy (SVM) for individual and 89.2% accuracy (KNN) and 89.3% accuracy (SVM) for generalized models. The accuracy difference between the two models ('individualized' and 'generalized') exists because stress impacts differently on every individual. Thus, a model derived from a population of individuals cannot be expected to be better than a model adjusted to the individual.

Since the typical human response to events is multimodal in nature, choosing multimodality in emotion research is well thought out [76]. All sorts of studies require collecting data, and physiological data must be as noise-ridden as possible to allow algorithms to operate more accurately [76]. In this work, information from both the physiological and behavioral responses of computer users was extracted. In order to develop a multimodal model as reliable and feasible for stress classification, modalities within the physiological response were combined, including HR, respiration, and EMG. The *Latent* [16] tool was used to record a person's behavior on the computer in order to access the behavioral component.

4

Materials and Methods

The current chapter provides an overview of the materials and procedures utilized to fulfill the work's aims. It contains an explanation of the computational methods applied, the extraction of the RR from accelerometer data, physiological signals and *Latent* [16] pre-processing techniques, and the features and parameters chosen for the models that were used. Furthermore, the data collection procedure and the protocols' development and implementation are also presented.

4.1 Computational Methods

The Python programming language was utilized in the code editor PyCharm to process and analyze the physiological signals as well as develop ML models. To extract the physiological features, several libraries were employed, including Pandas [77], NumPy [78], SciPy [79], TSFEL [80], and Biosignalsnotebooks [81]. Scikit-learn [82] was chosen as the package to create the ML models. The interactive protocols used in the pilot studies were implemented using the coding editor Brackets with the languages HTML and JavaScript, and the CSS framework Bootstrap. During these studies, the data collection tool *Latent* [16] recorded the user's activity on a browser page.

4.2 Extraction of Respiration Rates from Accelerometer Data

Previous research studies were able to extract respiration data from a smartphone placed on a person's chest, as indicated in Section 2.2. An algorithm was designed in order to use the RR derived from ACC data on the development of the stress detection classifier.

4.2.1 Preliminary Study

The PrevOccupAI project developed an android based smartphone application that allows evaluation of a person's biomechanics and assesses his/her posture using several different equipment, as illustrated in Figure 4.1. This application was used to collect the

data required for the development of the algorithm. The application begins by requesting the intended duration of the acquisition, followed by the selection of which equipment will be used. Once the equipment is selected, the acquisition procedure is ready to start. After initiating, the equipment will attempt to connect to the application. As soon as the pairing is established (when all left dots from the right-most image change from red to green), the recording process begins. The following images that illustrate how the devices are placed were provided by courtesy of the PrevOccupAI project.

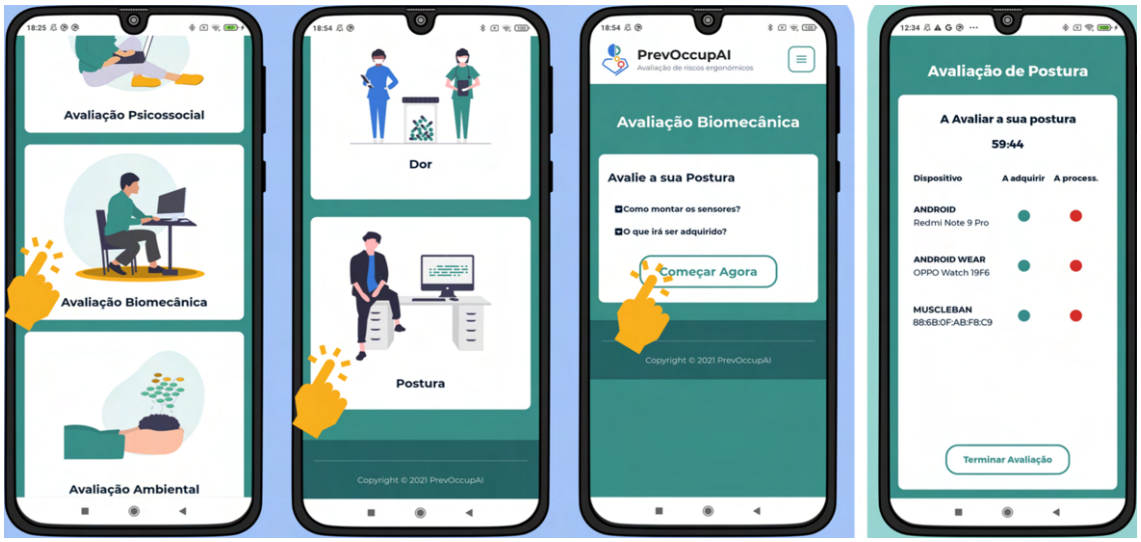


Figure 4.1: Example of the steps to do a biomechanical assessment. Retrieved from [83].

Figure 4.2: Smartphone placement.

For the preliminary study, a smartphone and a PLUX RIP band were chosen. The used smartphone was a Redmi Note 9 equipped with an ACC, gyroscope, magnetometer, noise-recorder and rotation vector. The ACC has a sample acquisition frequency of 100Hz. With the aid of a harness, the smartphone was mounted vertically on the subject's chest

as shown in Figure 4.2. The variation in the ACC induced by chest movements was used to obtain the individual's respiration signal. While breathing, this device would measure the acceleration components acc_x, acc_y and acc_z (with gravity acceleration) along the orthogonal x, y, and z directions, correspondingly. Figure 4.3 displays a visualization of the smartphone's axes

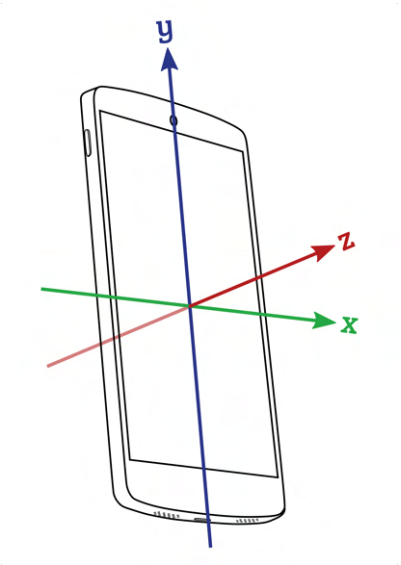


Figure 4.3: Smartphone Accelerometer's axes.

The RIP band from PLUX is a sensor that detects overall thoracic or abdominal displacement, making it more resistant to motion-induced artifacts. It contains an elastic strap adjustable in length (as shown in Figure 4.4), allowing the sensor to be used regardless of body structure. One of the many applications of this sensor is the monitoring of RR. In addition to the displacement signal, this equipment, like the smartphone, has an ACC for each one of the three axes. All signals are acquired at sample frequency of 1000Hz. In the PLUX RIP's datasheet [84], the raw data provided by this sensor is converted to displacement using equation 4.1. The Analog-to-digital converter (ADC) is the digital value sampled from the channel and n is the number of bits of the channel. The RIP's default resolution is 16-bit.

$$\text{RIP}(\%) = \frac{\text{ADC}}{2^n} - \frac{1}{2} \times 100\% \tag{4.1}$$

To validate the algorithm's reliability, a minor study was carried out utilizing these two sensors simultaneously. The study was performed on a 22 year-old female student, lasted for 40 minutes and began with a jumping period for further synchronization of the equipment. To cover different respiration patterns, three periods were distinguished in the study: a period when the subject was only concentrating on breathing, a period of forced fast breathing, and a period of breathing while mildly stressed. When the subject was focused on breathing, eyes were closed and the breathing pattern was as comfortable

as possible. For the fast breathing periods, the respiration rhythm was increased drastically, which is why these periods were only done three times over the 40-minute study and lasted no more than a minute. To simulate a bit of stress, an online high-difficulty level arithmetic exercise was carried out. Calculations would appear and had to be completed in 10 seconds. This activity was chosen to be as similar as possible to the one employed later in the stress-inducing protocol.

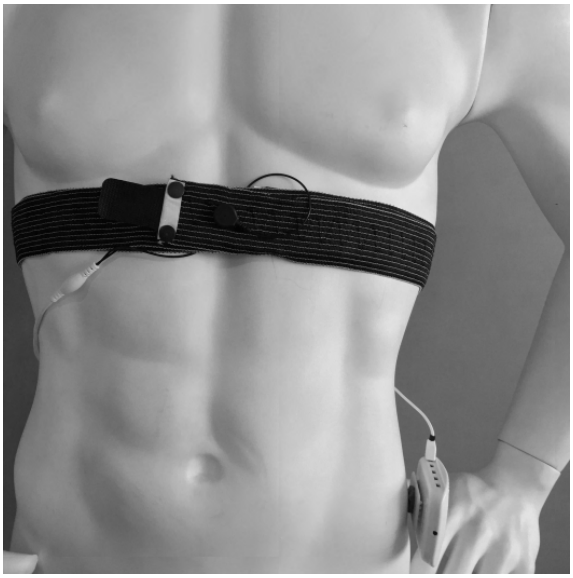


Figure 4.4: Inductive Respiration sensor placement. Retrieved from [84].

4.2.2 Algorithm Development

To develop this algorithm, the signals obtained from the two sensors were first synchronized, followed by the application of a forth-order low-pass butter-worth filter. Then, to reduce spurious peaks induced by motion, a noise threshold was applied to the peak and valley detection.

The degree of resemblance between a time series and a lagged version of another time series can be measured using cross correlation. Figure 4.5 shows the three types of possible lag. If the lag is negative, the red signal is delayed in reference to the black signal, and if the lag is positive, the process is reversed. If the lag is zero, the signals are perfectly synchronized. In this image, the highest correlation value corresponds to a positive lag value.

Synchronization of the sensors was done based on the jumping motion that was executed at the beginning of the study using a cross-correlation function from the Biosignalnotebooks library [81]. The jumping motion leaves a clearly distinguishable pattern on the ACC's y axis of each device (smartphone and RIP). Thus, the cross-correlation function was applied to the ACC's y-component of the devices. Because all of the RIP signals have a sampling frequency ten times superior to that of the smartphone ACC's signals,

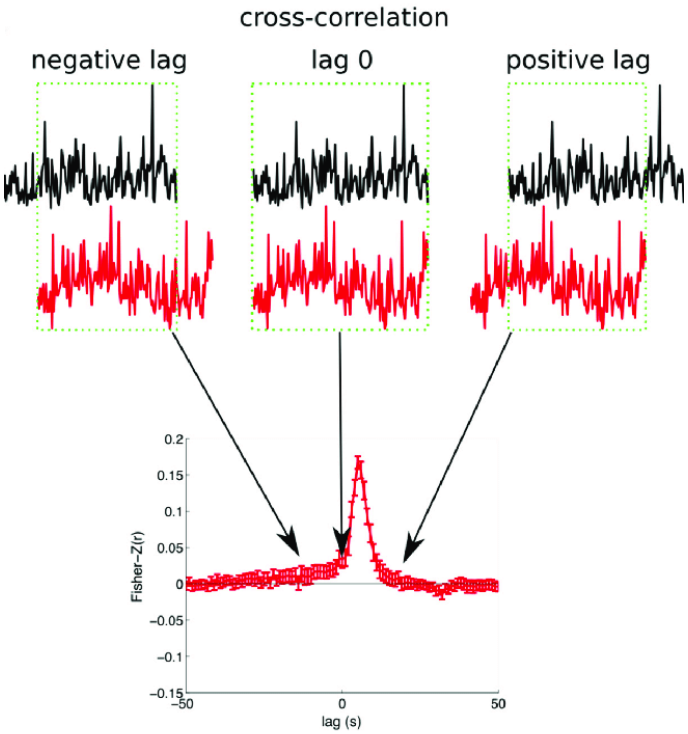


Figure 4.5: Cross Correlation. Retrieved from [85].

the synchronization was performed by implementation of a down-sample function. For every 10 values in the RIP ACC's y-component, the first one was chosen.

The most significant problems in ACC-based RR measurements are motion artifact removal and false peak detection [31]. To remove high frequency noise from the sum vector of the three axes' acceleration data (calculation showed on Equation 4.2), a third order low-pass butter-worth filter with a cutoff frequency of 0.5 Hz was used. The y axis, the z axis, a combination of these two, and the sum of the three axes were evaluated to decide which ones would be applied to detect the variation of the ACC that most accurately matched diaphragm movement. The sum vector was preferred over single axis data and y and z combination data, through experimental results.

$$ACC_{Total} = ACC_x^2 + ACC_y^2 + ACC_z^2 \tag{4.2}$$

The expansion and contraction of the diaphragm captured by the smartphone ACC appears as peaks and valleys in the ACC waveform. To overcome the possibility of false peak detection, an algorithm inspired on a research [31] was developed. After filtering the signal, peak and valley detection displayed in Figure 4.6 is performed in a 60-second window. Some of these peaks and valleys are caused by activities other than breathing. In order to remove this motion data, a noise threshold was implemented. Instead of being a stipulated value, this noise threshold was calculated for each specific

value. An empirically-determined 5-second sliding window was used to calculate the noise threshold value.

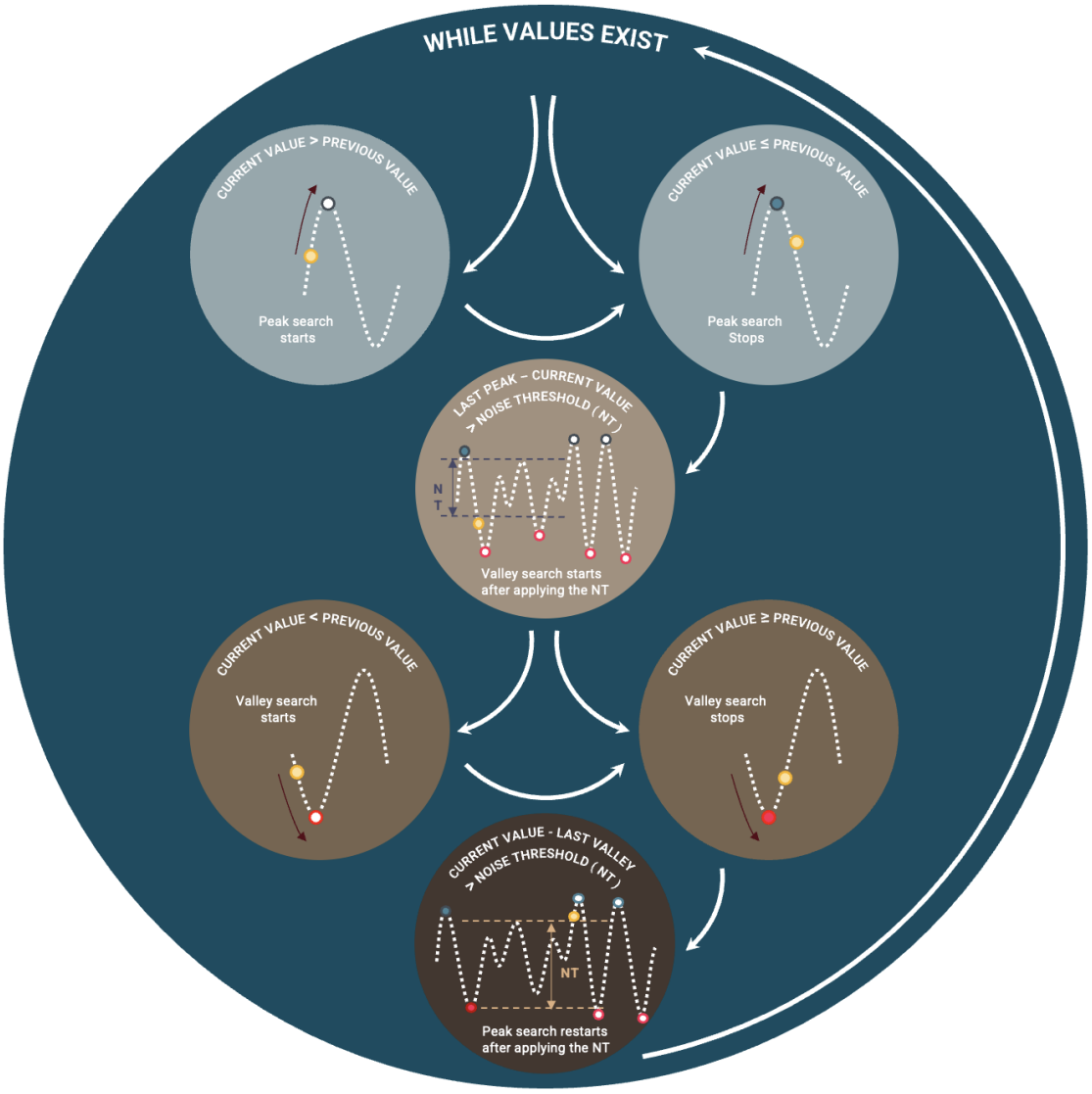


Figure 4.6: Schematic representation of respiration algorithm.

This window would compare previous and subsequent values to the current value. The highest and lowest values in this section are then subtracted from each other to give the maximum peak to valley amplitude of that period. All values less than half of that amplitude were considered noise. After all spurious peaks are removed, equation 4.3 is used to calculate the RR. The number of samples that exists between two peaks is averaged and divided by the sample frequency (f_s) in one minute periods. This returns an estimate of the rate at which an individual breathed during that minute.

$$\text{RR} = \frac{f_s}{\text{Peak to Peak sample count}} \times 60 \tag{4.3}$$

4.3 Experimental Protocols for Stress Detection

Two studies were carried out in order to collect the data required for the development of a stress detection classification model. For both protocols, the acquisition setup was kept the same. Both studies were carried out with the same subjects, in the same facilities and had a similar duration. The protocols were conducted in a quiet space with minimal outside noise. Participants, while sitting in a chair behind a desk, were able to observe the protocol interface on a laptop screen. The only allowed tool of interaction was the laptop's own trackpad. The computer sound was adjusted to a desired volume and was not changed during the studies. The layout of the two protocols was also as similar as possible despite their specificities. The purpose of the stress-inducing protocol was to create moments of cognitive and emotional stress in computer users whereas the relaxation-inducing protocol aimed to help participants to relax, lowering the chance of them experiencing a strong sudden emotion. Similar to other studies [28, 74, 86], the data extracted from the Stress-Inducing Protocol was labelled as "Stressed" whereas the one retrieved from the Relaxation-Inducing Protocol was labelled as "Not Stressed".

Physiological signals, such as HR, respiration (derived from ACC) and EMG were acquired from the 12 healthy volunteers (6 male and 6 female) that participated in the studies. The participant's ages were on average 25.75 ± 7.19 years and all of them were right-handed. In regards to professional occupation, half of the participants were students while the remaining half were workers. The participants were warned not to take any drugs or medication that might have an impact on their psychological state/awareness at least 24 hours before the experiment. They were also asked for their written informed consent, to fill out their demographic data and the PSS questionnaire [87], that quantifies a person's perceived stress experienced during the past month. This document is used in many studies as an objective stress marker [6, 17, 63, 65, 88]. Questions in the PSS are significant to see how life events are perceived. Depending on perception, two people who experience the same traumatic event can get different scores [87]. Since two studies were conducted, the PSS questionnaire was completed both times. It would not be accurate to infer that experienced stress in the previous month had remained the same [6].

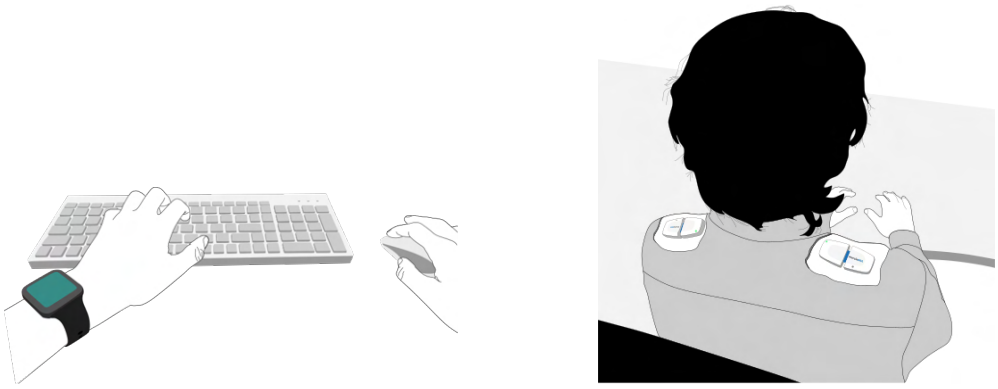
4.3.1 Acquisition Setup

4.3.1.1 Sensor Placement and Configuration

High-quality data is required to be accurate, complete, timely, relevant, adequately detailed and appropriately depicted. Instruments for measuring stress should be non-intrusive and discrete, since subjects may become more stressed as a result of invasive equipment [6]. Most of these requirements are met by the wearable physiological monitoring equipment outlined below. For the purpose of this thesis, a smartphone, a smartwatch and two muscleBANs were chosen. The muscleBANs are sensors that acquire muscle activity.

The Oppo smartwatch has the same sensors as the smartphone except for the noise-recorder. Instead, the watch can obtain the HR signal as explained in Section 2.1, with a 1Hz frequency. However, the watch has a limitation: it only collects data for about one minute out of every three, and if its battery is weak, it might not even comply with this rule. During the protocols, the smartwatch will be positioned on the wrist corresponding to the participant's least-dominant side (illustration on the left side of Figure 4.7) because any wrist movement can significantly distort the PPG signal [26]. The participant group did not contain any left-handed individuals.

The EMG signals are acquired at a 1000Hz frequency using two muscleBANs from PLUX. These devices contain, in addition to an EMG sensor, an ACC and a magnetometer. The muscleBANs are positioned on the upper trapezoids in accordance with SENIAM recommended placement [89] as displayed on the right side of Figure 4.7. To lower the electrode-skin impedance and achieve a better fixation of the electrodes, proper skin preparation is required [89]. Therefore, skin cleansing with alcohol was done before the electrodes were applied.



(a) Smartwatch. (b) MuscleBANs.

Figure 4.7: Smartwatch and MuscleBANs placement.

4.3.1.2 Latent Setup

As stated in Section 1.2, *Latent* [16] is a data collection tool that is integrated as an extension into the Google Chrome web browser to track users' activities while using it. Thus, the protocols had to run on Chrome in order to extract data from the HCI during the studies. After concluding several steps, including the installation of Docker and MongoDB, *Latent* was ready to be used. When pressing the *Latent* extension icon, a window as illustrated in Figure 4.8 appears on the screen. The left-most icon, when selected, shows what parameters can be recorded. These parameters are mouse interaction, keyboard interaction, geolocation, browser tab screenshots, audio from microphone, video camera snapshots and Document Object Model (DOM) as shown in the Figure 4.9.

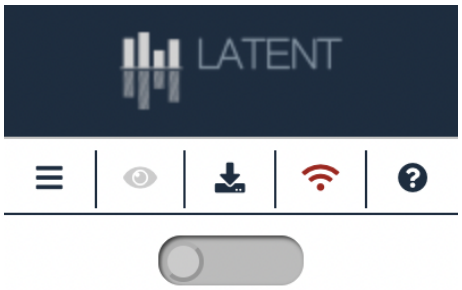


Figure 4.8: *Latent* web browser extension acquisition control. On top, icons from left to right: Open options page; turn on/off the acquisition preview; open the acquisition results page; checks if connection is being established with serve and "about" information. The bottom button is to start/stop the acquisitions. Retrieved from [16].

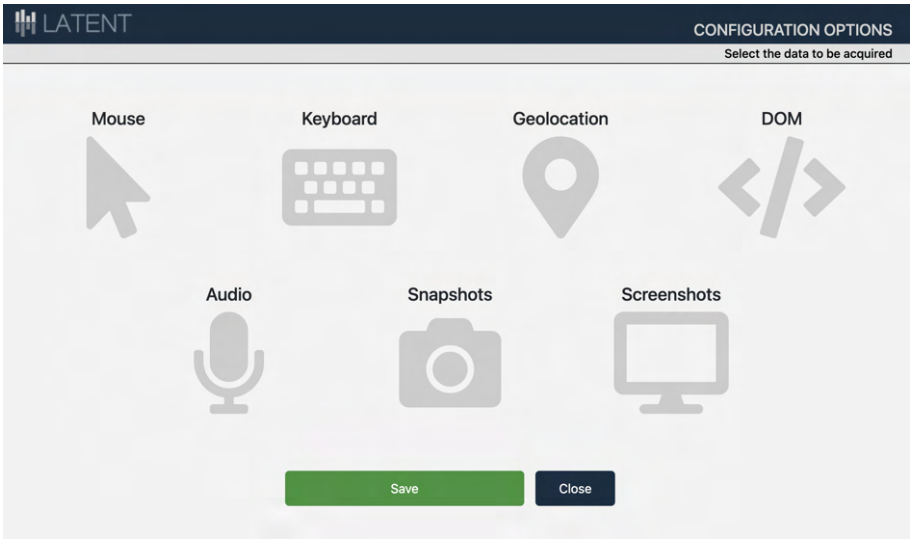


Figure 4.9: *Latent* open options page. Retrieved from [16].

During the acquisition protocols, screenshots of active browser tabs and mouse interaction were activated. The protocol was designed in a way that only a mouse is necessary for interaction, thus "Mouse" option was chosen in the *Latent* options menu. It provides the most pertinent data from user browsing behavior such as mouse position within the browser window, left and right button activity, and mouse wheel scrolling. The HTML element's XPath is also captured when the mouse is hovering over it. XPath is an abbreviation for XML Path Language and is mostly used in XSLT, although it can also be used to navigate through the DOM of any XML-like language, such as HTML [90]. The keyboard file, although it might have been interesting for HCI analysis, was not extracted because there was no need for keyboard interaction during the protocols. The screenshots option was activated to facilitate access of the competitive component in the mental calculations task.

4.3.2 Protocol Description

As mentioned in Section 4.1, protocols were developed in the Brackets code editor with the languages HTML and JavaScript, and the CSS framework Bootstrap.

As soon as the protocols started, the first requirement was to jump repeatedly for 10 seconds with arms extended along the body and eyes looking forward in order to enable future synchronization of all the collected files. The volunteers also had to perform a Reference Voluntary Contraction (RVC) for 20 seconds, which was extracted to later normalize the EMG amplitude. They were told to look straight ahead and raise their arms perpendicularly to the chest. Their elbows should be fully extended and wrists should be straight with palms facing down.

4.3.2.1 Stress-Inducing Protocol

Stress may be cognitive, emotional, or physically based [25]. Since computer work isn't a significant source of physical stress (there is no lifting of weights or the need for a lot of skeletal movement) this type of stress was not induced. Only one cognitive and one emotional task - respectively, the "Norinder Test" [91] and the "Sing a Song Test" [92] - were selected in order to keep the procedure brief and avoid bothering the participants with an overextensive protocol.

The "Norinder Test" is an arithmetic test that has to be performed under time constraints. This test is used to induce mental stress in studies [65–67]. The implementation of the "Norinder Test" for this protocol, similar to another study [66], is based on finishing 27 calculations within a 2:30 minutes time frame. The test had 2 timers with a colored circle changing from green to red gradually. The first timer displayed the total amount of time available to complete the task. The second timer had the time the participant could take to complete each calculation. When this timer reached 5 seconds, a stopwatch sound started to play, alerting the players that they didn't have much time left to select an answer. There were four possible options to be selected. When the correct one was

picked, a "success sound" would beep, and for two seconds, the correct answer would be displayed in green and the incorrect answers would be displayed in red as illustrated in Figure 4.10. A "wrong calculation" buzzer sounded when a mistake was made, and a totally red screen with the word "Wrong" appeared, blocking the page for three seconds. The subjects then received an additional 10 seconds to find out the answer once more. However, if they didn't pick an answer before the end of the 10-second countdown, the calculation would change, and they would no longer be able to respond to that particular one.

In the end of all acquisitions, an e-mail was sent with the final scores in a global ranking with all the participants, who had previously been informed that a small prize would be awarded to the best ranked person.

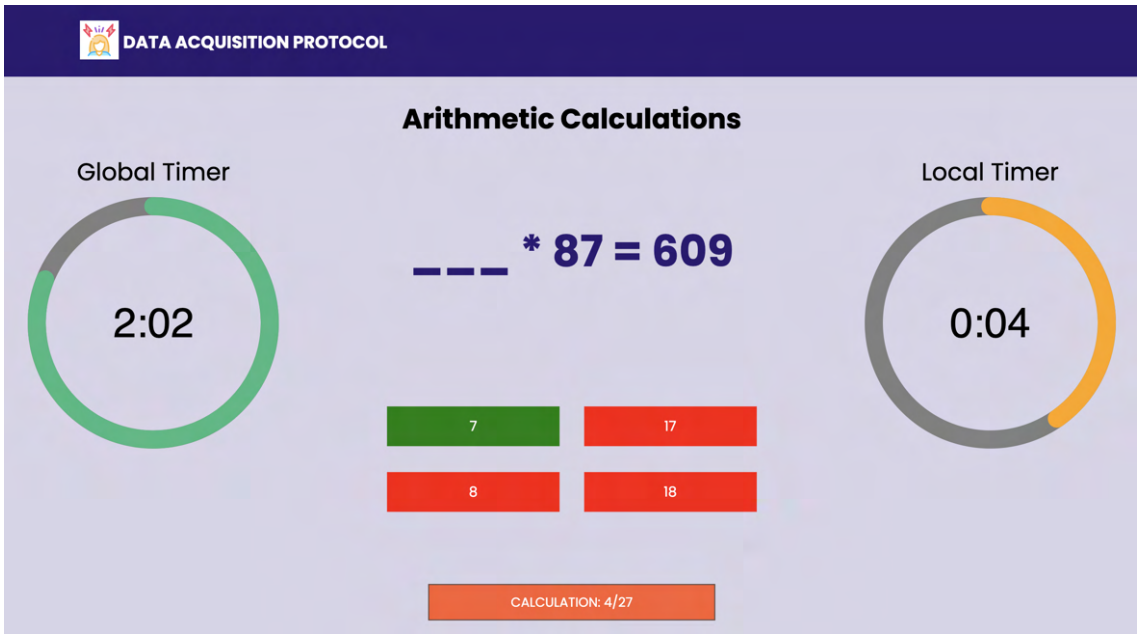


Figure 4.10: Right answer selected in the cognitive task.

To induce emotional stress, the "Sing a Song Test" was chosen. Used by several studies [6, 92, 93], this test aims to induce emotional stress in an ethical way that does not result in long-term detrimental effects on subjects, unlike the Montreal Imaging Stress Task or the International Affective Picture System, where some images can be so disturbing that the effects of stress can linger [6]. Participants were instructed to remain seated in front of the computer monitor and silently read 10 messages that would appear on the screen. This test also included a 10-second timer, with a colored circle gradually changing from green to red to indicate the duration each message would be displayed on the screen. It was indicated that one of the messages could contain a hidden task to be carried out after the subsequent timer reached 0. A stopwatch sound was also playing after 5 seconds had passed for each message. Because the messages could not arouse any stressful emotions, nine neutral sentences were selected with the tenth saying: "HIDDEN TASK: Think of

a song from your childhood. When the clock stops, sing the song out loud." (shown on Figure 4.11).

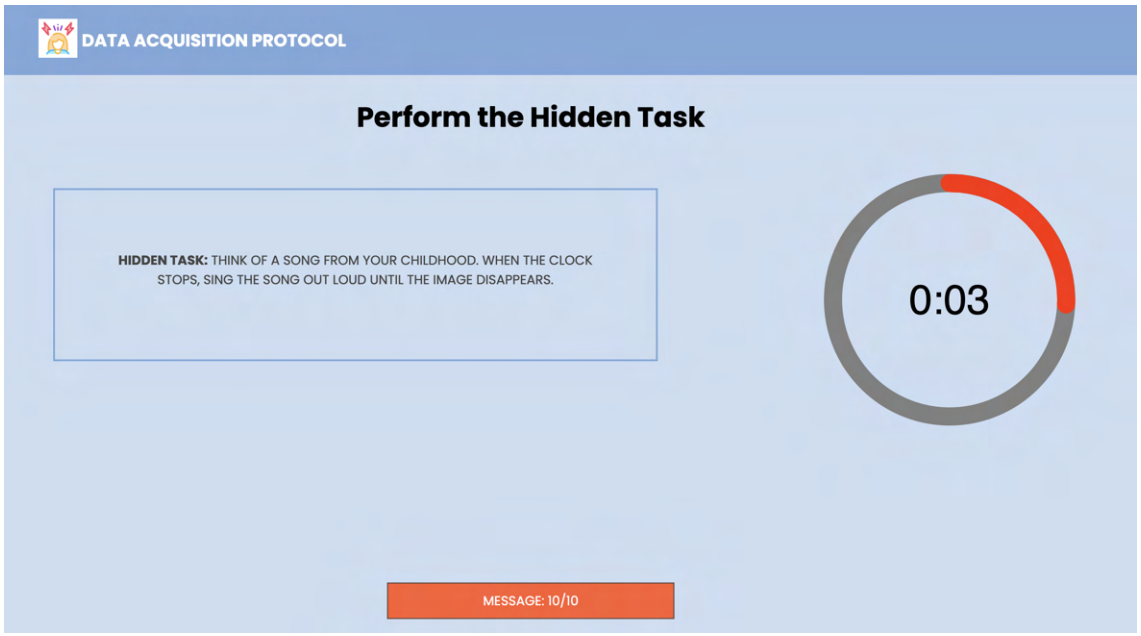


Figure 4.11: Hidden Task page in the emotional task.

4.3.2.2 Relaxation-Inducing Protocol

When designing the relaxation-inducing protocol, the tasks within the protocol needed to be developed in such a way that the computer user feels relaxed, but not to a point where he/she stops paying attention to the computer, as this usually does not happen in a normal work scenario. The best approach to ensure this is to establish a protocol in which some gentle relaxation is forced on the individual, as it is impossible to guarantee that the person is not stressed with other matters prior to the study. This was done in the first half of the protocol. The second part of the protocol tried to create an environment in which participants were attentive to the computer but did not experience any big emotion. To avoid evoking strong emotions, the entire protocol was created with a neutral design and was implemented in monochromatic gray-scale because it is the only color that has no direct psychological impact [94, 95]. Low-volume relaxing music was also played throughout the protocol to help set the mood [17].

Several relaxation methods incorporate the use of calm, deep breathing [96]. The 4-7-8 breathing technique was chosen because it has shown an effect on self-regulating stress in other studies [97, 98] and reduces both oxygen consumption and HR [99]. Participants were asked to sit with their back straight and softly press the tip of their tongue against the top of their mouth throughout the entire exercise. The exercise started by fully exhaling through the mouth. Then, the subjects inhaled slowly and quietly through their nose while mentally counting to four. Afterwards, they had to hold their breath for a count of

seven and perform an eight second-long exhale [100]. This cycle was repeated four times before they could move on to the next task. Figure 4.12 shows the a screenshot of the protocol when people were performing this task.

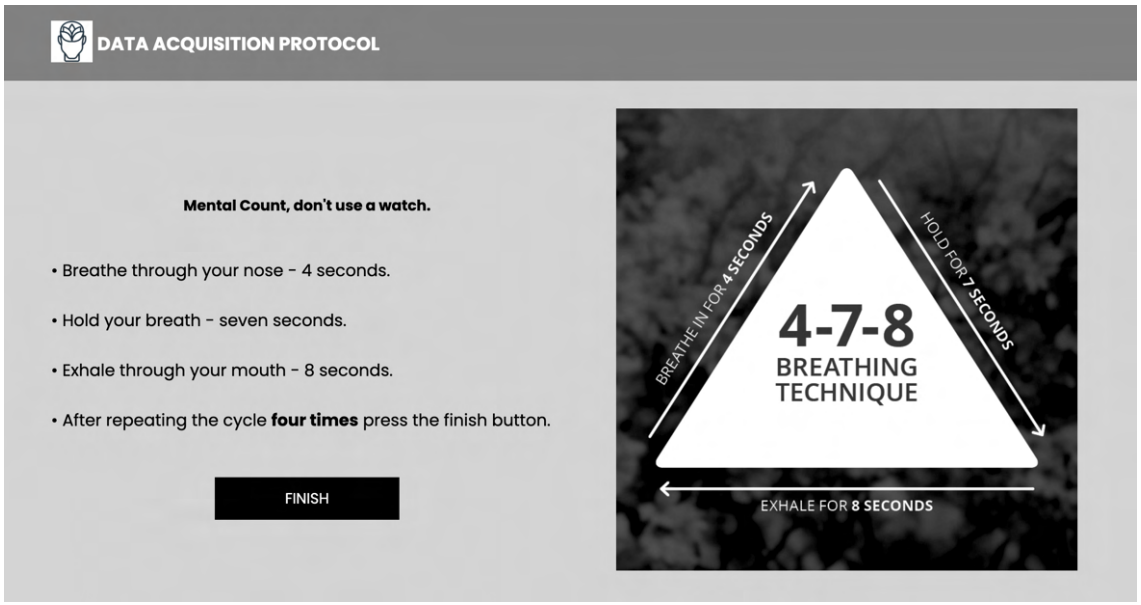


Figure 4.12: 4-7-8 breathing task.

After inducing relaxation, the subject can proceed to the second part of the protocol. The intent of this part was to maintain the person's state of relaxation after the breathing exercise while still keeping them focused on the computer, creating a close simulation of a calm computer user. To achieve that, a picture and a sentence tasks with 12 images and 12 sentences, respectively, were presented to the subject.

First, participants were instructed to remain as relaxed as possible while enjoying the images that would pass on the screen. There was no need for computer interaction in this task. Nature and fractal images in shades of green and blue were presented. Natural landscapes are well known for their calming effects [95]. Green is an emotionally tranquil color, while blue encourages reflection and introspection [95]. Soft blues soothe the mind and improve concentration [95]. Fractal installations that are evocative of nature can decrease occupant stress levels in addition to easing physical discomfort by alleviating the pressure that the environment's abnormal spatial frequencies place on the eyes and mind [101]. These patterns successfully reduce physiological arousal, stress levels, and help to capture attention [101, 102]. Figures 4.13 and 4.13 illustrate these two type of images present in the protocol.

After observing the pictures, the sentence task would begin. Similarly to the picture task, participants were told to stay as as relaxed as possible while reading the sentences on the computer monitor. Some neutral sentences inspired on a research [94] with small cognitive thought would appear on the screen as Figure 4.15 shows. Despite being entirely neutral because they only provide facts, these types of sentences compel the reader to

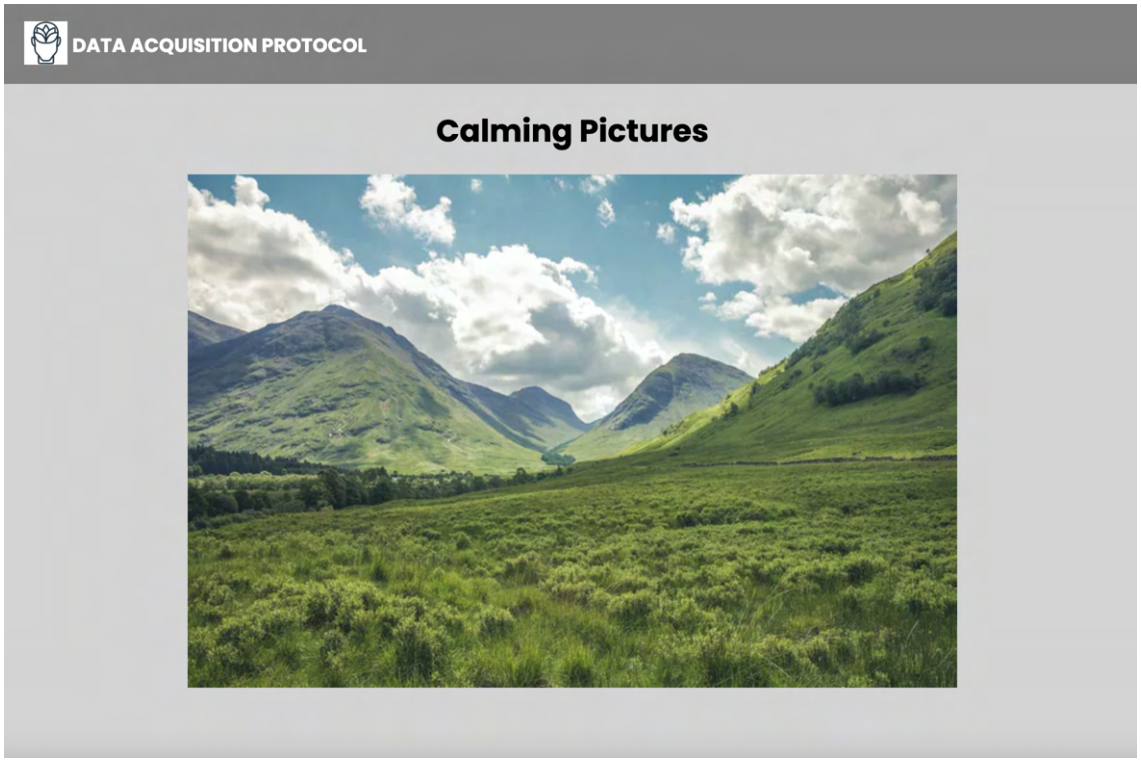


Figure 4.13: Natures Landscape image in Picture Task.

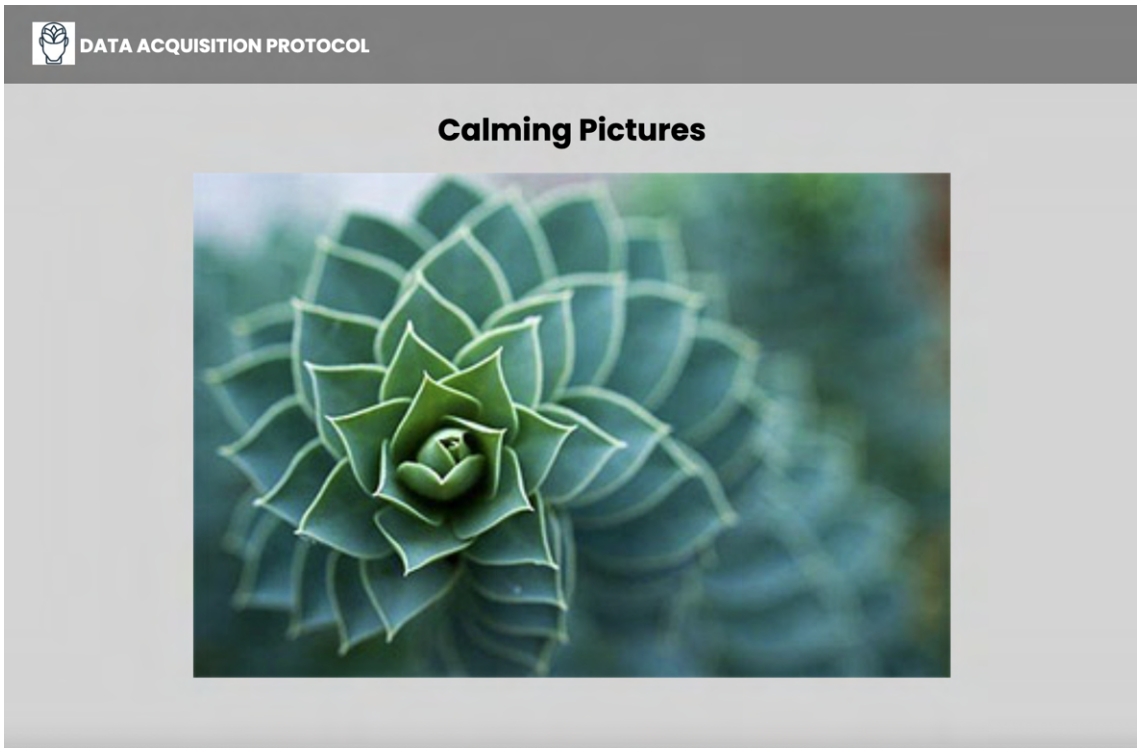
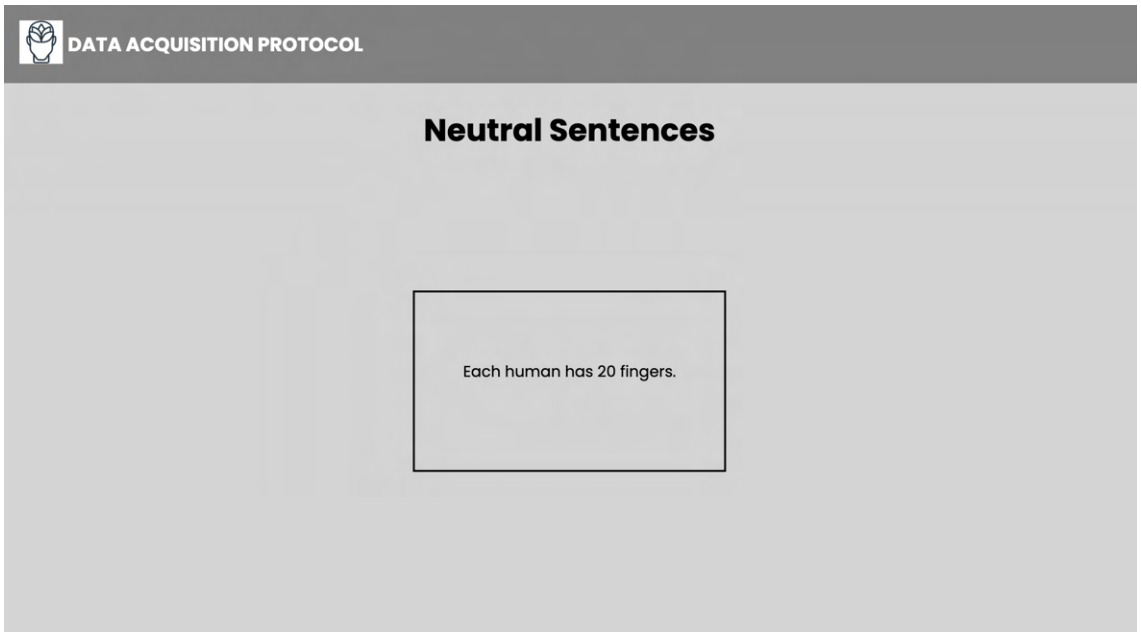


Figure 4.14: Fractal image in Picture Task.

verify whether or nor they were true, which resulted in unconscious cognitive thinking without arousing any strong emotion.



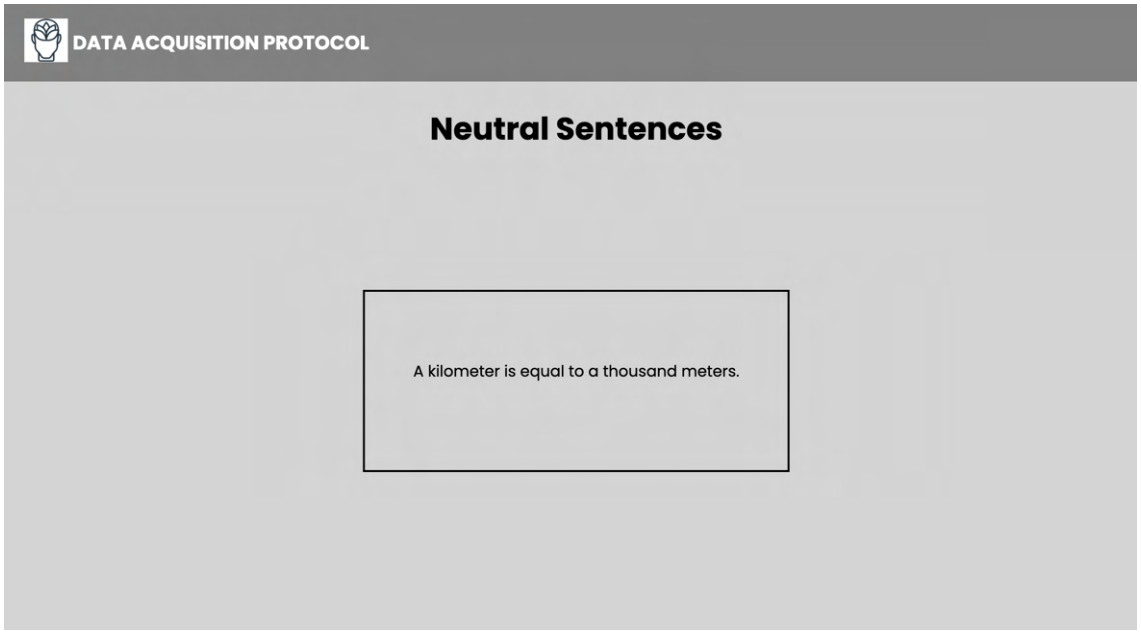


Figure 4.15: Sentences in "Sentence Task".

4.4 Pre-Processing and Feature Extraction

4.4.1 Data Synchronization

Since data is retrieved from four different devices there are some challenges regarding information merging. For all acquisition starting times to be in sync, the data must be synchronized using timestamps. Also, only data retrieved while subjects were performing tasks was considered. The parts of the protocol in which people jump, raise their arms or read what they had to do were eliminated because it could not be assured whether or not the participants were stressed in those activities.

4.4.1.1 Sensors

Both the smartphone and the smartwatch provided a file containing the timestamps associated with an x, y, and z acceleration as shown in Table 4.1. The muscleBANs file did not have a timestamp, but that information was added using the the eletromyography sensor's sample frequency.

Table 4.1: Section of smartphone obtained file.

Timestamp	xAcc	yAcc	zAcc
5223861213415	-0.62	9.452	1.129
5223871237400	-0.38	9.813	1.164
5223881261467	-0.368	9.679	1.136
5223891285481	-0.584	9.416	1.081
5223901309457	-0.483	9.624	1.052
5223911333490	-0.327	9.76	1.09
5223921357721	-0.526	9.528	1.124
5223931381746	-0.653	9.49	1.097
5223941405808	-0.445	9.756	1.126
5223951429823	-0.397	9.698	1.124
5223961454029	-0.596	9.456	1.081
5223971477972	-0.586	9.574	1.083
5223981502068	-0.38	9.789	1.081
5223991526063	-0.471	9.602	1.107
5224001550044	-0.672	9.411	1.126
5224011574288	-0.564	9.622	1.085
5224021598255	-0.368	9.794	1.047
5224031622317	-0.502	9.6	1.054
5224041646317	-0.655	9.459	1.109
...	...	...	...

Similarly to the study described in 4.2.1, the sensor synchronization was done based on the initial jumping period using the cross-correlation function from the Biosignalnotebooks library [81]. Thus, the cross-correlation function was also applied to the y-axis of the ACC's devices. The smartphone and smartwatch were synchronized first followed by

the smartphone and muscleBANs. Since the smartphone has the application where the acquisition process starts, it was always the first device to begin data acquisition. As a result, all lags resulting from the cross-correlation were positive. Figure 4.16 shows four graphs showing the syncronization process before and after. On the top-left side, a graph with all axis from the smartphone and muscleBAN before syncronization is displayed, and on the bottom-left side, only the ACC's y axis is presented. The smartphone and muscleBAN time series after synchronization are shown at the bottom. One can see that the muscleBAN time-series shifted to the right to be synchronized with the smartphone, resulting on a positive lag.

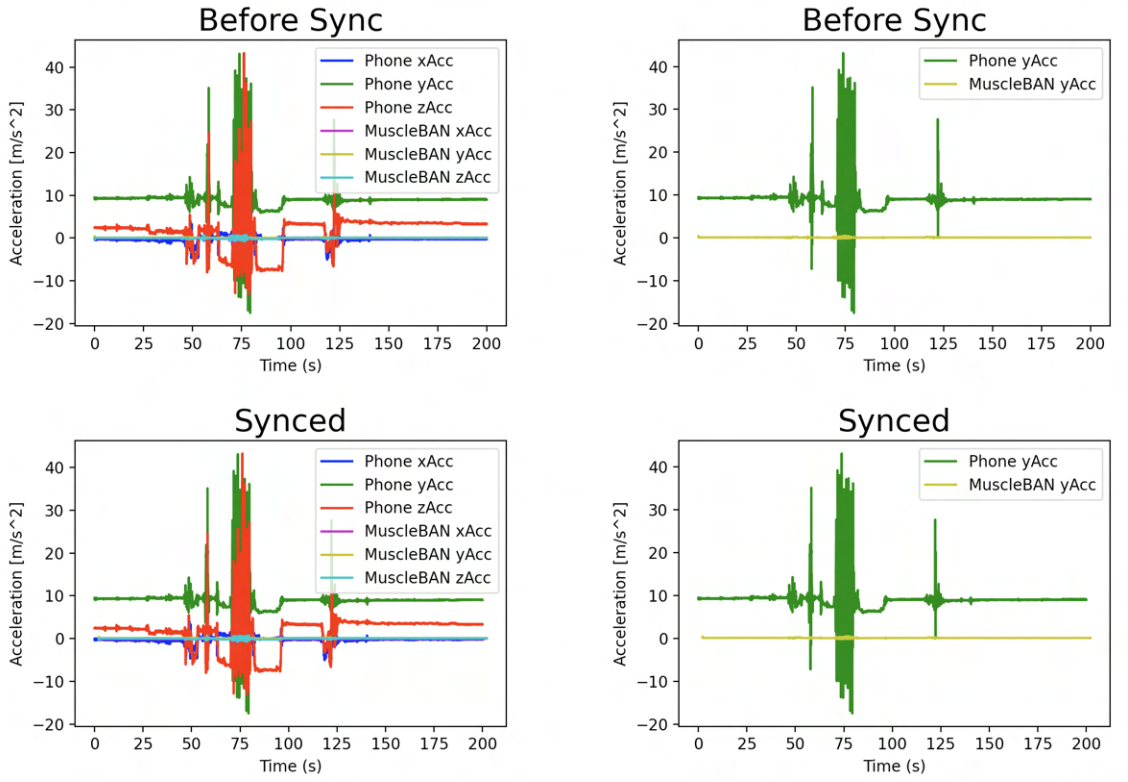


Figure 4.16: On top, from left to right: all smartphone and muscleBAN accelerometer's axis before synchronization; only the smartphone and muscleBAN accelerometer's y axis before synchronisation. On bottom, from left to right: all smartphone and muscleBAN accelerometer's axis after synchronization; only the smartphone and muscleBAN accelerometer's y axis after synchronization.

4.4.1.2 Latent

The *Latent* [16] data had to be synchronized with the remaining physiological signals as well. Every time a participant executed a "left click", it would appear written in the file on a column labeled "mouse interactions". Left clicks occurred anytime participants pressed a button that transported them to another HTML page (except for the calculations task). When people were ready to begin the protocol, they could press the start button

on the first page of the procedure. The second page detailed how the jumping period should be carried out. After pressing the "Jump" button, a third page appeared with a picture displayed for 10 seconds, equivalent to the time they had to jump. Following that, a page describing how the RVC should be performed was shown. Therefore, after three left mouse clicks, all accelerometers displayed a "plateau" on the z-component corresponding to people curving their bodies to read how the RVC should be performed. People remained standing after jumping since RVC would also be completed that way, and they would bend over to press the button that started the 20-second countdown in which they would be performing the RVC.

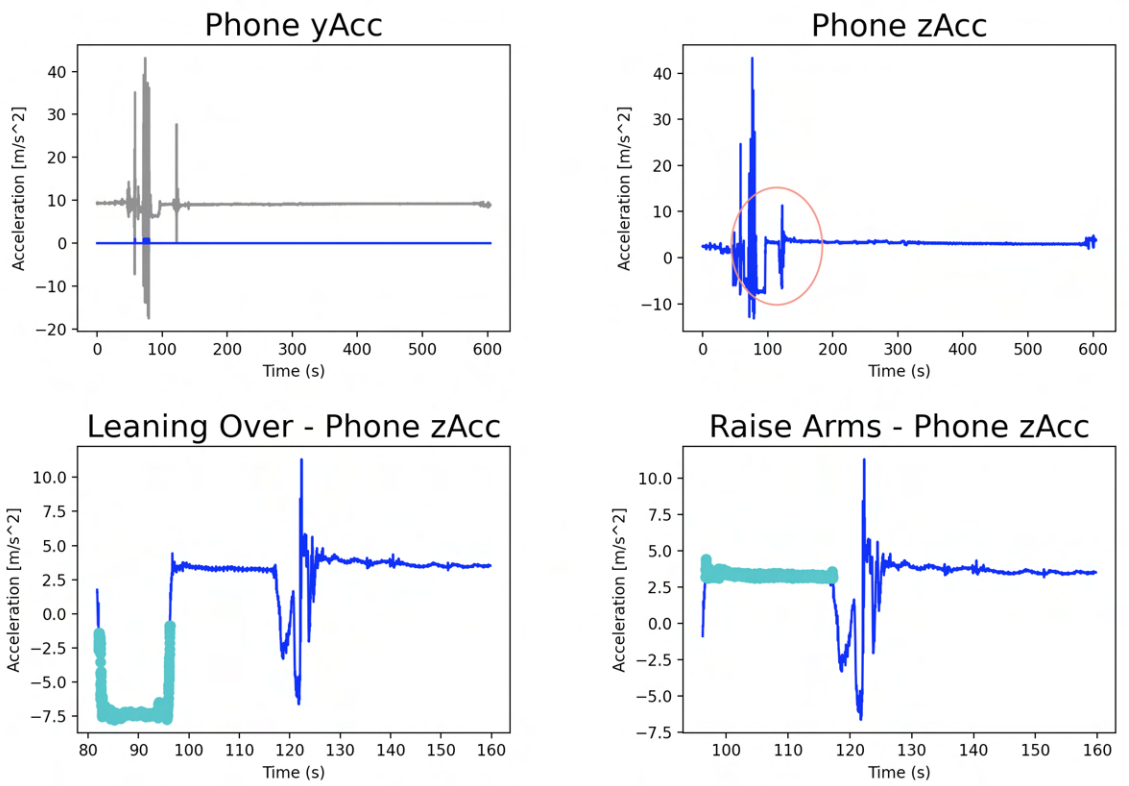


Figure 4.17: On top, from left to right: detection of jumping period; original acceleration on z-axis. On bottom, from left to right: detection of the "plateau" on the accelerometer's z-component; detection of the raise arms period on the accelerometer's z-component.

Figure 4.17 displays the detection of the jumping period on the top left. This detection was achieved using a function that calculated if the difference between the current acceleration value and the previous one was greater than 30m/s^2 (experimentally determined value), and if it was, then the current value was part of the jumping period. On the top-right, the original complete z-axis acceleration is displayed and the pink circle corresponds to the leaning over and arm raising periods. On the bottom-left, the light blue dots represent the acceleration values when the participants were leaning over. For the development of this algorithm, an 80-second time window was chosen. This time period corresponded to the maximum time found among all participants between the

end of the jumping and the end of the arms raising task. Then, the average acceleration of that time period was calculated. When the current acceleration value was higher than the value obtained before, all values under the average were deemed to be part of the "plateau". If seven seconds (empirically determined value) elapsed between the current value and the previous one, the leaning period could be considered to be over. Finally, at the bottom right, the "plateau" corresponding to the arm raising period is shown with its respective acceleration values also colored in light blue. To identify these values, the same algorithm used to detect the leaning over values was employed. For an acceleration value to be considered part of the raising arms period, it had to be above the average this time, and the search would only begin when the current acceleration value was lower than the previous value. If two seconds (experimentally determined value) passed between the current and preceding values, the arms raising period was over.

After understanding how the differentiation between the jumping, leaning over, and arm raising periods was achieved, it is simpler to comprehend how the synchronization between the *Latent* and the smartphone was done. The file synchronization was accomplished by replacing the *Latent* timestamp in which the participants pressed the mouse for the third time with the smartphone timestamp where the minimum value within the "plateau" was encountered. Since the data from the jumping and raising arms periods won't be used for the classification model, there will be no problem synchronizing the *Latent* file only after those events. Following this logic of using clicks, the timestamps where the cognitive and emotional tasks happened can now be identified in order to remove other irrelevant moments of the protocol, such as reading the tasks.

4.4.2 Data Processing and Feature Extraction

After all biosignals have been collected and synchronized, they must be processed in order to extract the most information out of them. To achieve this, an overlapping window function was created. The window size and overlap duration were dependent on each signal.

4.4.2.1 Heart Rate

For the purpose of this thesis, a continuous HR signal would be ideal so that more features could be extracted from it. However, such endeavour was not possible. As stated above, the smartwatch performs a one minute data collection every three minutes. To overcome this issue, a function that eliminated the missing time intervals was implemented. If at least 25 seconds passed between two subsequent time values (the minimum amount of time that passed between two acquisitions), all previous values were saved, and new time values were searched complying to the requirements above. Additionally, not all of the data acquired by the watch was used to extract features; just the one regarding the performance of the cognitive and emotional tasks was used. Hence, in order to gain some more information for future feature extraction, the HR signal was re-sampled

from 1Hz to 5Hz for each acquisition period. Then, the sliding window function was also applied. The signal was segmented into 8-second windows. This value was chosen because it was the shortest smartwatch acquisition time found among participants. Thus, it was possible to guarantee that in every acquisition some sort of feature extraction could be done. Half of this value was chosen as the window overlap value, in order to achieve some significant information gain.

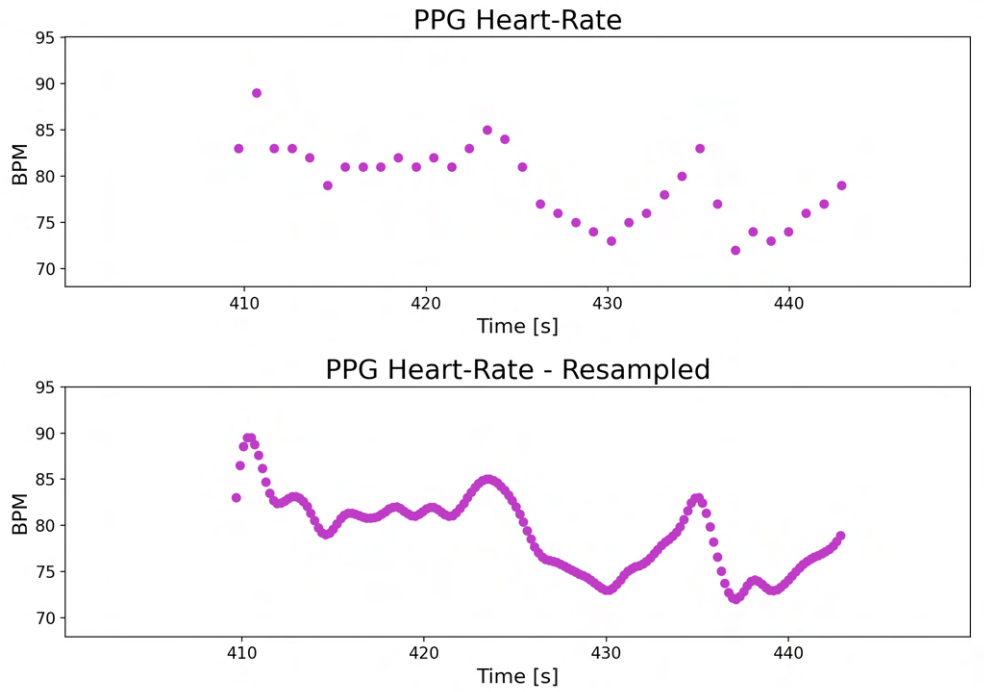


Figure 4.18: Top: Photoplethysmography heart rate acquisition over time. Bottom: Photoplethysmography heart rate acquisition re-sampled over time.

The smartwatch extracts data in BPM; this data can be transformed into a RR interval estimate. By using Equation 4.4 — which divides the 6000 milliseconds that make up a minute by the number of beats it contains — it can be estimated what the interval between two R-peaks is in milliseconds.

$$\text{RR Interval} = \frac{6000}{\text{BPM}} \tag{4.4}$$

The RR interval can be used to extract three different types of features: time, frequency, and non-linear features. As was done in other studies, this thesis sought to extract temporal and frequency features since these were the most promising domains for the identification of stressful moments, particularly time-related ones [3, 28, 29, 58, 103, 104]. Tables 4.2 and 4.3 list the extracted features along with the formula and brief descriptions of each.

Table 4.2: Time-Domain extracted HR features.

Feature	Unit	Formula	Description
mRR	ms	$\frac{\sum_{i=1}^{N}(RR_i)}{N}$	RR Interval Mean
maxHR	bpm	$max(BPM)$	HR interval max
mHR	bpm	$\frac{\sum_{i=1}^{N}(BPM)}{N}$	HR Interval Mean
minHR	bpm	$min(BPM)$	HR Interval Min
SDRR	ms	$\sqrt{\frac{\sum_{i=1}^{N}(RR_i - mRR)^2}{N-1}}$	RR Standard Deviation
SDHR	bpm	$\sqrt{\frac{\sum_{i=1}^{N}(BPM - mHR)^2}{N-1}}$	HR Standard Deviation
CVRR		$\frac{SDRR \times 100}{mRR}$	RR Coefficient of Variance
RMSSD	ms	$\sqrt{(mean((RR_{i+1} - RR_i)^2)}$	Root Mean Square of Successive RR differences
RR20		$Count(\lvert RR_{i+1} - RR_i \rvert)_{>20ms}$	Number of Pairs of adjacent RR Intervals differing by more than 20 ms to all RR intervals
PRR20	%	$\frac{Count(\lvert RR_{i+1} - RR_i \rvert)_{>20ms} x100}{\text{N-1}}$	Percentage of Number of Pairs of adjacent RR intervals differing by more than 20 ms to all RR intervals
RR50		$Count(\lvert RR_{i+1} - RR_i \rvert)_{>50ms}$	Number of Pairs of adjacent RR intervals differing by more than 50 ms to all RR intervals
PRR50		$\frac{Count(\lvert RR_{i+1} - RR_i \rvert)_{>50ms} x100}{N-1}$	Percentage of Number of Pairs of adjacent RR intervals differing by more than 50 ms to all RR intervals

Table 4.3: Frequency-Domain extracted HR features.

Feature	Unit	Formula	Description
VLF Power	ms^2	Power spectrum from 0.003 to 0.04 Hz	Power Spectrum of Very Low Frequency
LF Power	ms^2	Power spectrum from 0.04 to 0.15 Hz	Power Spectrum of Low Frequency
HF Power	ms^2	Power spectrum from 0.15 to 0.4 Hz	Power Spectrum of High Frequency
Total Power	ms^2	VLF+LF+HF	Total Power Spectrum of Frequency
nVLF	%	$\frac{\text{VLF}}{(\text{VLF+LF+HF})} \times 100$	Normalized Very Low Frequency Spectrum
nLF	%	$\frac{\text{LF}}{(\text{VLF+LF+HF})} \times 100$	Normalized Low Frequency Spectrum
nHF	%	$\frac{\text{HF}}{(\text{VLF+LF+HF})} \times 100$	Normalized High Frequency Spectrum
dLHF	%	|NLF-NHF|	Difference of Normalized Low Frequency Spectrum and Normalized High Frequency Spectrum
nuLF		$\frac{\text{LF}}{(\text{LF+HF})} \times 100$	Low Frequency in Normalized Unites
nuHF		$\frac{\text{HF}}{(\text{LF+HF})} \times 100$	High Frequency in Normalized Unites
SMI		$\frac{\text{LF}}{(\text{LF+HF})}$	Symphathetic Modulation Index
VMI		$\frac{\text{HF}}{(\text{LF+HF})}$	Vagal Modulation Index
SVI		$\frac{\text{LF}}{\text{HF}}$	Symphatovagal Balance Index

4.4.2.2 Respiration

The respiration signal was obtained using the smartphone's ACC. First, the total acceleration was calculated, then a fourth-order low-pass butter-worth filter with a cutoff frequency of 0.5Hz (experimentally determined value) was applied. The sliding overlapping function was then used. The chosen window size was 60-second, in agreement to how the RR is calculated, and the overlapping time was 10 seconds (experimentally determined value). Finally, the RR was calculated recurring to the algorithm explained in Section 4.2.2. Because the RR was not the only metric available for extraction from the ACC signal, some other features mentioned on Table 4.4. Also, since these features were computed using TSFEL's library [80], unlike the tables presented for the HR, this table will not contain information about the formulas used to obtain them.

Table 4.4: Respiration extracted features.

TSFEL Features	Description
Nr. Peaks	Number of Peaks
RR	Respiration Rate
MAD	Median Absolute Deviation
STD	Standard Deviation
Var	Variance
AE	Absolute Energy
Area	Area under the curve
PP Distance	Peak-to-Peak Distance

4.4.2.3 EMG

The EMG signal was the one which provided the most information, for the reason that the muscleBANs acquired at a higher frequency than the other sensors. According to the PLUX muscleBANs' datasheet [105], the raw EMG data is converted to mV using equation 4.5. Like the RIP's sensor, the muscleBANs' default resolution is also 16-bit. The VCC and the G_{EMG} are also used for this conversion. The first is the operation voltage and the second is the sensor gain (2500V and 1100, respectively for the muscleBANs device).

$$EMG_{mV} = \frac{\frac{\text{ADC}}{2^n} - \frac{1}{2} \times \text{VCC}}{G_{EMG}} \tag{4.5}$$

For the EMG signal, features were extracted in different pre-processing steps. This was done because some features needed signal properties that were eliminated by certain pre-processing algorithms. Figure 4.19 shows the entire pre-processing and feature extraction workflow and Figure 4.20 aids in graphically illustrating the impact that each process step has on the signal.

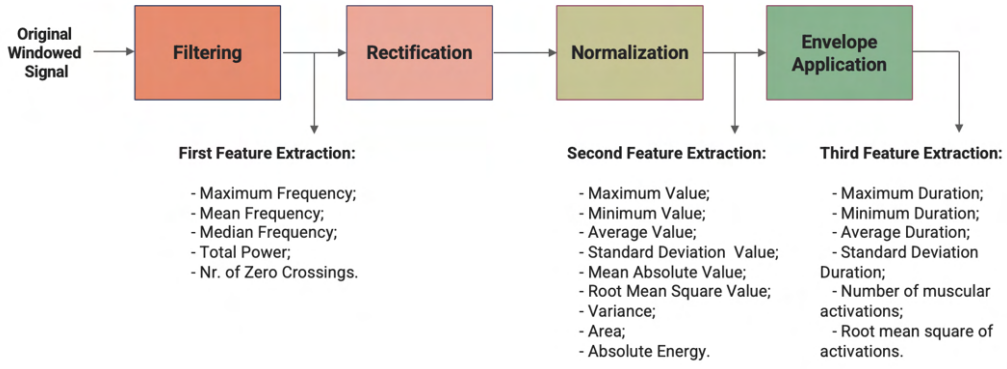


Figure 4.19: Diagram with the pre-processing steps and the stages of feature extraction.

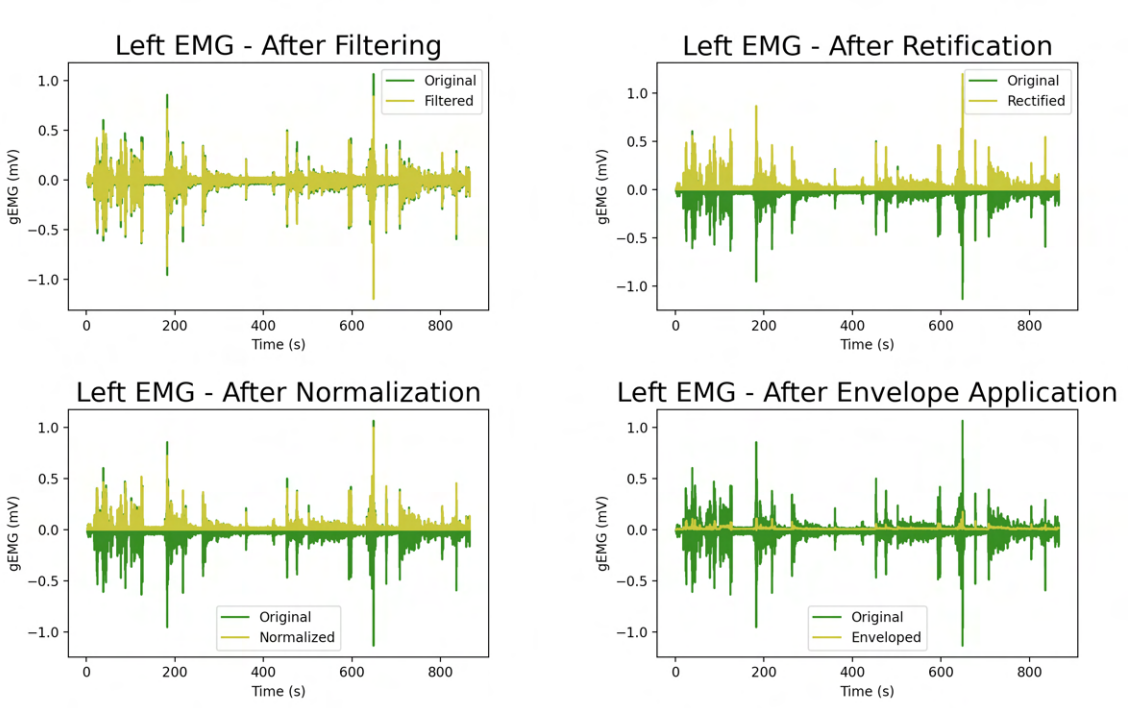


Figure 4.20: Pre-processing of electromyography signal.

To keep consistency with the other biosignals, after converting it to mV, the EMG signal was segmented into windows of size 60 seconds with an overlap of 10 seconds (empirically determined value). Then, the signals were filtered using a a fourth-order band-pass butter-worth filter. The window size and the order of the filter were chosen in

accordance to a research [17] that also used left and right trapezius EMG for stress detection. To choose the cutoff frequencies, a power spectral density of the signal was plotted, as shown in Figure 4.21, and it was determined that the most common frequencies in the signal were mainly between 30Hz and 310Hz. After filtering, the first set of 4 frequency features and 1 statistical were extracted. These were the maximum, mean and median frequency; the total power and the number of zero crossings.

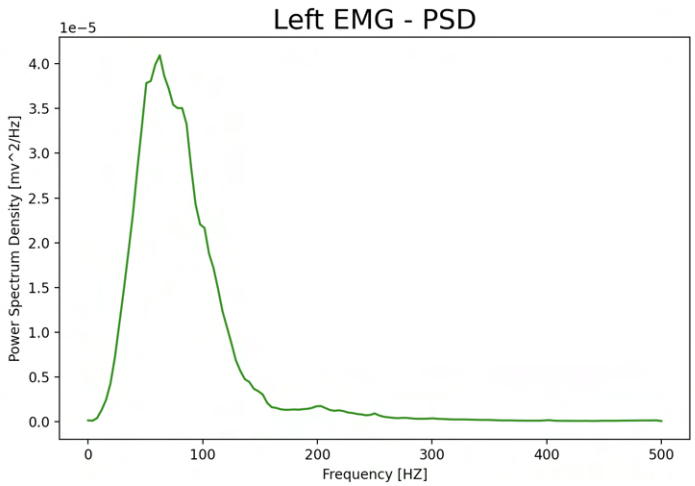


Figure 4.21: Power spectrum density of the left trapezius electromyography signal.

Subsequently, the signal was rectified (replacing all values with their absolute value) followed by a normalization using a maximum norm scheme. The original intention was to normalize the EMG signal using the RVC value obtained during the protocol. Since obtaining a Maximal Voluntary Contraction (MVC) was impractical, a RVC was done. In a normalization process, the amplitude of the EMG signal is converted into a scaled number, typically the percentage of a MVC or RVC from a certain task. In this case, the RVC was supposed to reflect the maximum EMG result. Because there were many amplitudes that were greater than the RVC value, the normalization process was changed. Instead, the maximum signal amplitude was used as a scaling factor for the entire signal. After these steps, the next set of 9 statistical features were extracted from the windowed signal, consisting of maximum, minimum, average, standard deviation, mean absolute and root mean square values as well as variance, absolute energy and area.

In a final step, a fourth-order low-pass butter-worth filter with a cutoff frequency of 2Hz (empirically determined value) was utilized to make an envelope allowing for clear detection of muscular activity periods from the signal. From this, a final set of 6 features were extracted. These features focused on the muscular activation periods and consisted of the maximum, minimum, average and standard deviation duration; the root mean square and the number of muscular activations. Overall, this means that a total of 20 features were extracted from the EMG signals. Table 4.5 presents an overview of all features and at which stage they were extracted.

Table 4.5: EMG extracted features.

Feature	Unit	Formula	Description
MNF	Hz	$mean(frequency)$	Mean Frequency
MDF	Hz	$median(frequency)$	Median Frequency
MAXF	Hz	$max(frequency)$	Maximum Frequency
Area	ms^2	$\int_0^{T_{max}} \mathrm{E}\ \mathrm{dt}$	Area
Total Power	ms^2	$\int_0^{F_{max}} P_x\ \mathrm{df}$	Total Power Spectrum of Frequency
ZC		$\{x_i > 0 \text{and } x_{i+1} < 0\}$ or $\{x_i < 0 \text{ and} x_{i+1} > 0\}$ and $\lvert x_i - x_{i+1}\rvert \geq \epsilon$	Zero Crossing
MAXV	mV	$max(E)$	Maximum Value
MINV	mV	$min(E)$	Minimum Value
ANV	mV	$avg(E)$	Average Value
STDV	mV	$std(E)$	Standard Deviation Value
RMSE	mV	$\sqrt{\frac{1}{N}\sum_{i=1}^{N} E^2}$	Root Mean Square Value
Var	mV	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(E-\overline{E})^2}$	Variance
Energy	W	$\sum_{i=1}^{N}(\lvert E\rvert)^2$	Energy
RMSA	mV	$\sqrt{\frac{1}{N}\sum_{i=1}^{N} A^2}$	Root Mean Square of Activation
NA		$count(A)$	Number of Muscular Activations
MAV	mV	$\sqrt{\frac{1}{N}\lvert E\rvert}$	Mean Absolute Value
MAXD	ms	$max(A_{Duration})$	Maximum Duration
MIND	ms	$min(A_{Duration})$	Minimum Duration
AND	ms	$mean(A_{Duration})$	Average Duration
STDD	ms	$std(A_{Duration})$	Standard Deviation Duration

4.4.2.4 Multimodal Windowing Scheme

The multimodal model receives features extracted from all acquired signals: the HR from the smartwatch, the respiration derived from the smartphone's ACC, and the EMG signals from the two muscleBANs. As the acquisition periods for these sensors differ significantly, only periods in which all sensors acquire at the same time were considered for feature extraction. Figure 4.22 illustrates the windowing scheme applied for feature extraction. For simplicity, only one EMG sensor is shown. The periods at which all sensors acquire at the same time are highlighted in gray. The window sizes for each signal were kept the same as those for the individual models, that being 8 seconds for the HR and 60 seconds for the respiration and EMG respectively. It was ensured that the end of each window was aligned in time. All windows were shifted by 4 seconds until the end of the HR data was reached (end of a gray section).

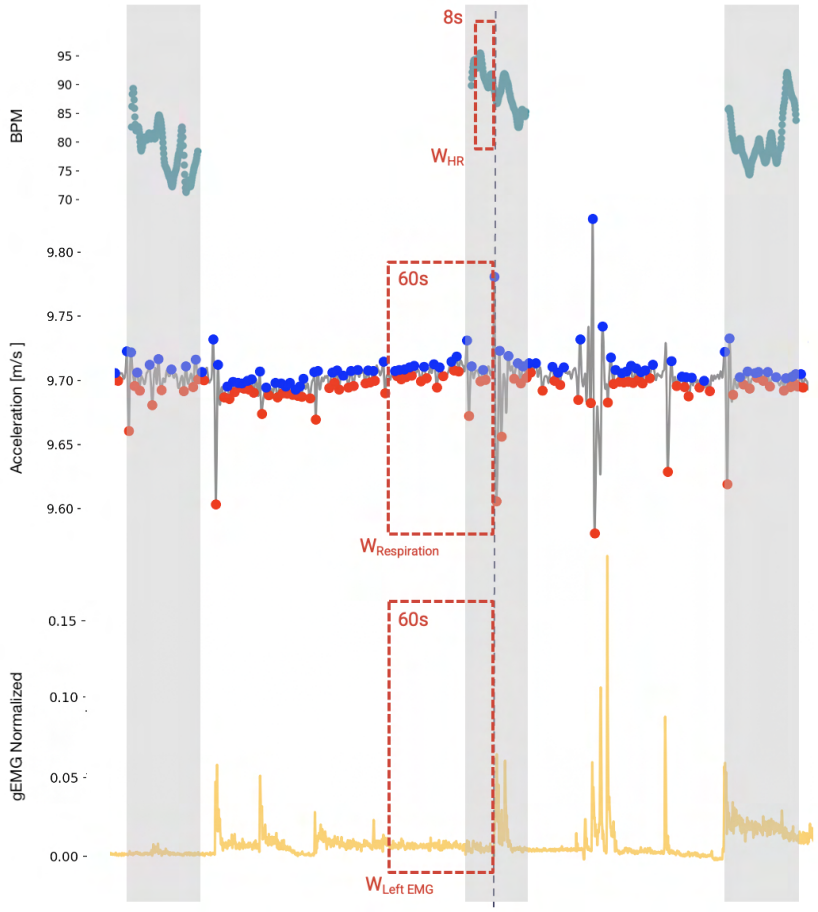


Figure 4.22: Multimodal Windowing Scheme.

4.4.2.5 Latent

Although it was part of the thesis plan, feature extraction form the data collected through *Latent* [16] was not further pursued. Several circumstances led to this decision. Firstly, the installation of *Latent* took more time than initially expected. Throughout the installation process, several errors that had to be resolved, such as incompatibilities with the operating system, were encountered. Resolving these issues took several days for a single computer. Thus, the setup explained in Section 4.3.1.2 was used, in which all participants used the same computer during data acquisition. Secondly, after exploring the *Latent* data, it became clear that the participants' computer usage, during the protocols, was too minimal (e.g., not enough interaction with the mouse and none with the keyboard). Thus, it was concluded that the data would not be representative enough to extract any behavioral features for classification.

4.5 Machine Learning Algorithms

Several models were analyzed with the three ML algorithms described in Section 2.5: HR, respiration, left EMG, right EMG, sum of left and right EMG and a combination of all three biosignals.

To assess a model's performance, it must first be trained before being tested. To accomplish this, the labeled data-set was randomly separated into a training and a testing group of samples. Two data splits were done, one using 75% of training samples and 25% of testing (default split used by the Scikit-learn [82] employed function), and another one with 60% of training and 40% of testing (to determine if the classification remained the same or if it deteriorated as a result of having less training data). Both data split were done in a stratified way so that each class was as equally represented as possible inside the training and testing set. Because models were always tested with unknown samples it ensured that they could be implemented in real-world situations to unknown data, avoiding overfitting. Since linear models (like the employed linear SVM) produce distinct outcomes depending on whether the data is normalized or not, both the training and testing set were normalized. To obtain the metrics for each classifier and for each model, the algorithm was tested 5 times with different test sets. Accuracy is the metric that predicts how many total instances were well classified. However, evaluating accuracy alone can be misleading. For example, a classifier that correctly classifies half the time may just be classifying all samples in the same class if there is an equal distribution of classes. In this way, the other metrics presented in Section 2.5 must also be evaluated together with the models.

4.5.1 Feature Selection

The contribution of each feature to the classifier's training is displayed in Tables 4.7 and 4.6. For feature selection, first, correlated features were removed using the TSFEL library [80]. Then the Recursive Feature Elimination (RFE) function supplied by the Scikit-learn library [82] was applied to remove the weakest feature(s) until a specified number was reached. This function was run several times and the rank displayed on both tables corresponds to the order of the best-found features during the RF and SVM classifiers' training. The blue colored features are the ones that are commonly better between the RF and SVM and the red ones have a non-significant contribution to the classification. As explained in Section 2.5.3 the KNN classifier does not provide feature weights or coefficient attributes; instead, each prediction is exclusively dependent on the distance between two data points. As a result, and because feature selection also has an impact on this classifier's performance, the features that obtained the best results in the SVM and RF classifiers were used in KNN as well. The HR model was trained using the 4 out of the 7 extracted features colored in blue in Table 4.7. The respiration one used 7 features: the top 3 marked on the table plus the PSS, the RR, area under the curve and variance. The left and right EMG models used the first 12 features that appear in the table as well as the EMG total model which also used the first 20 features that appear in Table 4.6. Finally, the multimodal model used the blue-flagged features of the HR, respiration and left EMG making a total of 20 features.

Table 4.6: Electromyography selected features. Blue cells represent the best and red cells represent the under performing features for the Random Forest and the Support Vector Machine classifiers.

Rank	EMG		Rank	EMG		Rank	EMG	
	RF	SVM		RF	SVM		RF	SVM
1	**R: RMSE**	**R: Energy**	15	L: Energy	L: Min Value	29	L: Median Frequency	L: Std Duration
2	L: Nr. Zero Crossing	L: Max Value	16	R: Max Value	L: Median Frequency	30	R: Std Value	L: Nr. Muscular Activations
3	**R: Energy**	**R: Avg Value**	17	L: Avg Value	R: Avg Duration	31	L: Avg Duration	R: Std Duration
4	**R: RMSA**	**R: RMSE**	18	L: RMSE	L: Std Value	32	R: Max Duration	L: RMSE
5	**R: Nr. Zero Crossing**	**R: Area**	19	L: Total Power	L: RMSA	33	R: Nr. Muscular Activations	L: Avg Value
6	**R: Area**	L: Mean Frequency	20	L: MAV	R: Min Duration	34	R: Std Duration	R: Std Value
7	**R: Avg Value**	**R: Nr. Zero Crossing**	21	L: RMSA	R: Max Duration	35	L: Std Duration	R: Total Power
8	**R: Mean Frequency**	**R: MAV**	22	R: Min Value	R: Nr. Muscular Activations	36	R: Avg Duration	R: Variance
9	**R: MAV**	**R: Mean Frequency**	23	L: Min Value	L: Energy	37	L: Max Duration	PSS
10	L: Max Value	**R: Max Value**	24	R: Median Frequency	R: Max Frequency	38	R: Min Duration	L: Max Frequency
11	R: Nr. Muscular Activations	R: Median Frequency	25	L: Nr. Muscular Activations	L: Max Duration	39	L: Min Duration	L: Area
12	PSS	L: Variance	26	L: Std Value	R: RMSA	40	R: Max Frequency	L: MAV
13	R: Mean Frequency	R: Min Value	27	L: Variance	L: Nr. Zero Crossing	41	L: Max Frequency	L: Min Duration
14	L: Area	L: Total Power	28	R: Variance	L: Avg Duration			

Table 4.7: Heart rate, respiration and left and right electromyography selected features. Blue cells represent the best and red cells represent the under performing features for the Random Forest and the Support Vector Machine classifiers.

Rank	Heart-Rate		Respiration		Left EMG		Right EMG	
	RF	SVM	RF	SVM	RF	SVM	RF	SVM
1	mRR	mRR	Nr. Peaks	Nr. Peaks	Nr. Zero Crossing	Nr. Zero Crossing	RMSE	Energy
2	SDRR	SDRR	RR	RR	Max Value	Max Value	Nr. Zero Crossing	RMSE
3	CVRR	CVRR	Median absolute Deiation	Median absolute Deiation	Total Power	Mean Frequency	RMSA	Avg Value
4	PSS	PSS	Std Deviation	Std Deviation	PSS	Energy	Energy	Area
5	mHR	mHR	Variance	Variance	RMSE	Median Frequency	Avg Value	Nr. Zero Crossing
6	SDHR	SDHR	Absolute Energy	DAE8FCAbsolute Energy	Avg Value	RMSA	Area	MAV
7	CVRR	RMSSD	PSS	PSS	MAV	Avg Value	MAV	Max Value
8			Std Deviation	Std Deviation	Energy	Area	Total Power	Min Value
9			PP distance	PP distance	Mean Frequency	RMSE	Mean Frequency	Variance
10			Autocorrelation	Autocorrelation	Area	MAV	Max Value	Avg Duration
11					RMSA	Max Frequency	PSS	PSS
12					Min Value	Avg Duration	Nr. Muscular Activations	Std Value
13					Median Frequency	Nr. Muscular Activations	Median Frequency	Min Duration
14					Std Value	Max Duration	Min Value	Mean Frequency
15					Variance	PSS	Min Duration	RMSA
16					Std Duration	Std Duration	Avg Duration	Max Duration
17					Max Duration	Total Power	Std Duration	Max Frequency
18					Avg Duration	Std Value	Variance	Median Frequency
19					Nr. Muscular Activations	Variance	Max Duration	Std Duration
20					Min Duration	Min Value	Std Value	Total Power
21					Max Frequency	Min Duration	Max Frequency	Nr. Muscular Activations

4.5.2 Parameter Selection

All ML algorithms have specific hyperparameters that must be adjusted in order to improve model performance [106]. GridSearch was chosen to optimize the hyperparameters of the three different chosen algorithms. To choose the cross-validation splitting strategy for GridSearch, the Scikit-learn library RepeatedStratifiedKFold function [82] was used with 5 folds. The search for the best hyperparameters was done with a total of 15 iterations and with a validation size of 20%. Other iteration values (10, 15, and 20) were tried, but 15 was selected since it found the ideal balance between running a sufficient number of times for the findings to be useful while not consuming much time from the computer.

The hyperparameters that were tweaked per algorithm were the regularization parameter ("C") and the kernel coefficient for specific kernels ("Gamma") for the SVM classifier; the maximum depth of the tree ("Max Depth"), the number of features to consider when looking for the best split ("Max Features") and the number of trees in the forest ("Nr. Estimators") for the RF algorithm and the number of neighbors to use ("Neighbors Nr."), the weight function used in prediction ("Weights") and the distance metric to use for the tree ("Metric") for the KNN classifier. The best hyperparameters found with the Scikit-learn function [82] in each algorithm and for each model are presented in Tables 4.8, 4.9 and 4.10. With the exception of the RF's "Nr. Estimators" and the "Metric" and "Neighbors Nr." for the KNN algorithm, all model parameters remain constant. The HR and respiration models stand out using 6 and 2 neighbors, respectively, as opposed to the other models' 1 neighbor.

Table 4.8: Best Hyperparameters of all models for Support Vector Machine.

Classifier	**Signal**	**Best Parameters**		
		C	Gamma	Kernel
SVM	Heart Rate	10	10	Linear
	Respiration	10	10	Linear
	Left EMG	10	10	Linear
	Right EMG	10	10	Linear
	Total EMG	10	10	Linear
	Multimodal	10	10	Linear

Table 4.9: Best Hyperparameters of all models for Random Forest.

Classifier	**Signal**	**Best Parameters**		
		Max Depth	Max Features	Nr. Estimators
RF	Heart Rate	4	4	80
	Respiration	7	7	80
	Left EMG	12	12	80
	Right EMG	12	12	70
	Total EMG	20	20	90
	Multimodal	20	20	40

Table 4.10: Best Hyperparameters of all models for K-Nearest Neighbor.

Classifier	**Signal**	**Best Parameters**		
		Metric	Nr. Neighbors	Weights
KNN	Heart Rate	manhattan	6	Linear
	Respiration	manhattan	2	Linear
	Left EMG	manhattan	1	Linear
	Right EMG	manhattan	1	Linear
	Total EMG	manhattan	1	Linear
	Multimodal	manhattan	1	Linear

5

Results and Discussion

This Chapter presents and discusses all of the findings made throughout the development of this thesis. First, the developed respiration algorithm's reliability is evaluated. Then, the effects of the signals' pre-processing choices and feature selection are presented. The results and discussion regarding the ML models are also provided. Finally, several comparisons are done between the various developed models and those discovered in Chapter 3 to better assess their performance.

5.1 Respiration Rate Algorithm

For the development of the RR detection algorithm explained in Section 4.2 different ACC axes combinations were tested to see which one performed better. Research that utilized an ACC to extract respiration used the z-axis [35], or a combination of the three axes [31].

Figures 5.1, 5.2, and 5.3 show which peaks were detected by the algorithm based on comfortable slow breathing, fast breathing, and breathing while mildly stressed, respectively. The algorithm was run using y and z axes and two combinations of axes (yz and xyz). Combining more than one axis to detect diaphragm movements is an adequate method to add the various components of the respiration signal constructively and add the noise destructively. The respiration patterns extracted from the RIP sensor are shown as well for comparison.

Table 5.1 presents the number of peaks the algorithm has detected for each axis, combination of axes, and the RIP sensor. It is evident that for all the tested axes and combinations the number of detected peaks is close to the real number of peaks extracted from the RIP sensor. The z-axis was able to reach precisely the same number of peaks. A closer look at the slow breathing pattern in Figure 5.1 confirms that the peaks and valleys detected by the axes and the combinations correspond to those detected by the RIP sensor.

Table 5.1: Comparison of peaks and respiration rate detected by the different axes combinations and the respiration signal on a comfortable slow breathing period.

Axis	Nr. Peaks	Respiration Rate [BPM]
Y	45	8.1
Z	46	8.2
YZ	45	8.1
XYZ	45	8.1
Respiration	46	8.1

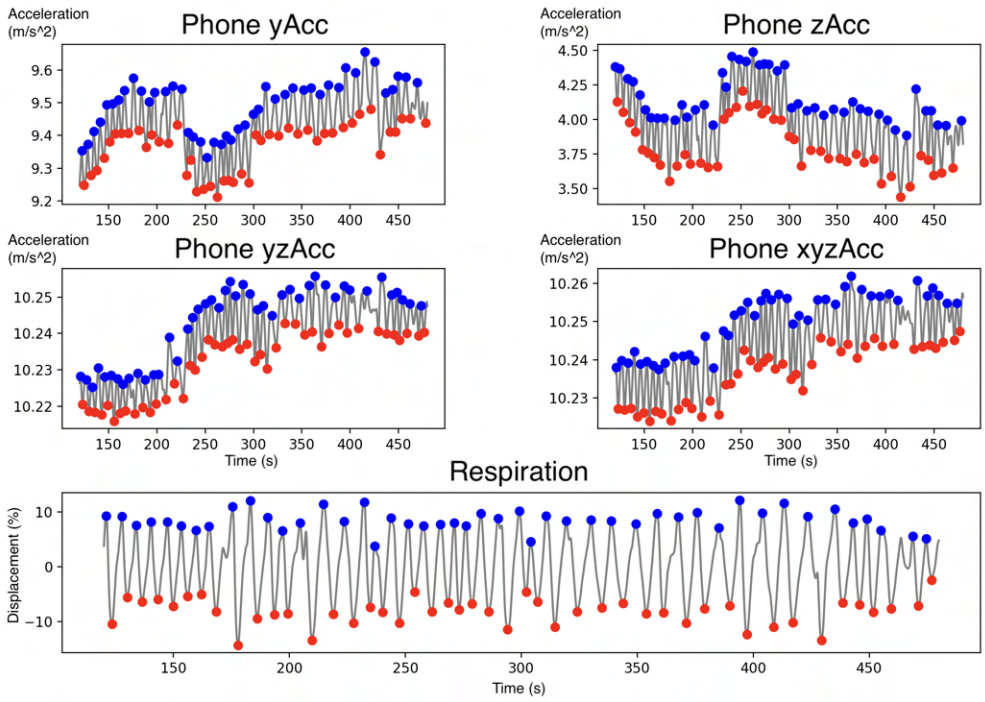


Figure 5.1: Graphical representation of the y-axis, z-axis, yz-combined-axes, xyz-combined-axes and respiration signal provided by the Inductive Respiration sensor on a comfortable slow breathing period.

For the fast breathing pattern (Figure 5.2) the expansion and contraction of the thoracic cavity are more evident, leading to more pronounced peaks and valleys, making the detection clearer. This observation can be verified by looking at Table 5.2 which reveals that all axes and combinations detect the same number of peaks, and the calculated RR rates are almost identical, with the z-axis attaining exactly the same RR value.

Table 5.2: Comparison of peaks and respiration rate detected by the different axes combinations and the respiration signal on a fast breathing period.

Axis	Nr. Peaks	Respiration Rate
Y	18	18.9
Z	18	18.4
YZ	18	18.7
XYZ	18	18.5
Respiration	18	18.4

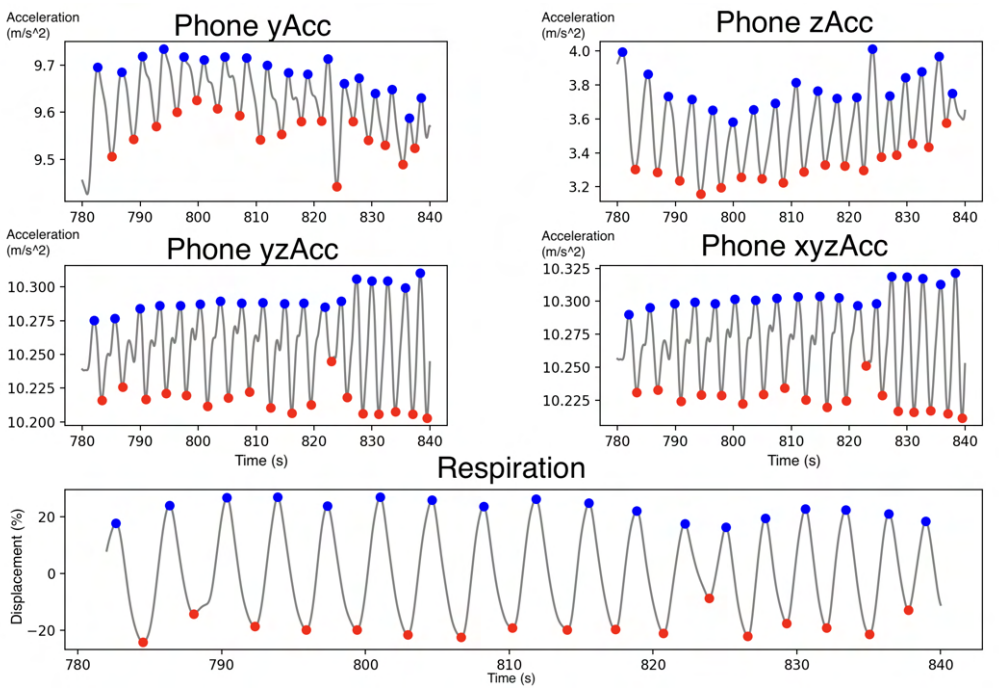


Figure 5.2: Graphical representation of the y-axis, z-axis, yz-combined-axes, xyz-combined-axes and respiration signal provided by the Inductive Respiration sensor on a fast breathing period.

When analyzing the algorithm results for the mild stress breathing pattern (Figure 5.3) the y-axis ACC component poorly captured the breathing pattern. The z-axis performed much better than the y-axis in peak detection and, consecutively, in estimating the respiration rate as shown in Table 5.3. The combination of the y and z axes proved to be the best choice for estimating the number of peaks and the RR in the period that most closely resembles the stress experienced by a worker when comparing to the RIP sensor followed by the combination of the three axes.

Table 5.3: Comparison of peaks and respiration rate detected by the different axes combinations and the respiration signal on a mild stressful period.

Axis	Nr. Peaks	Respiration Rate
Y	2	2.2
Z	43	22.8
YZ	45	23.4
XYZ	46	23.8
Respiration	44	23.3

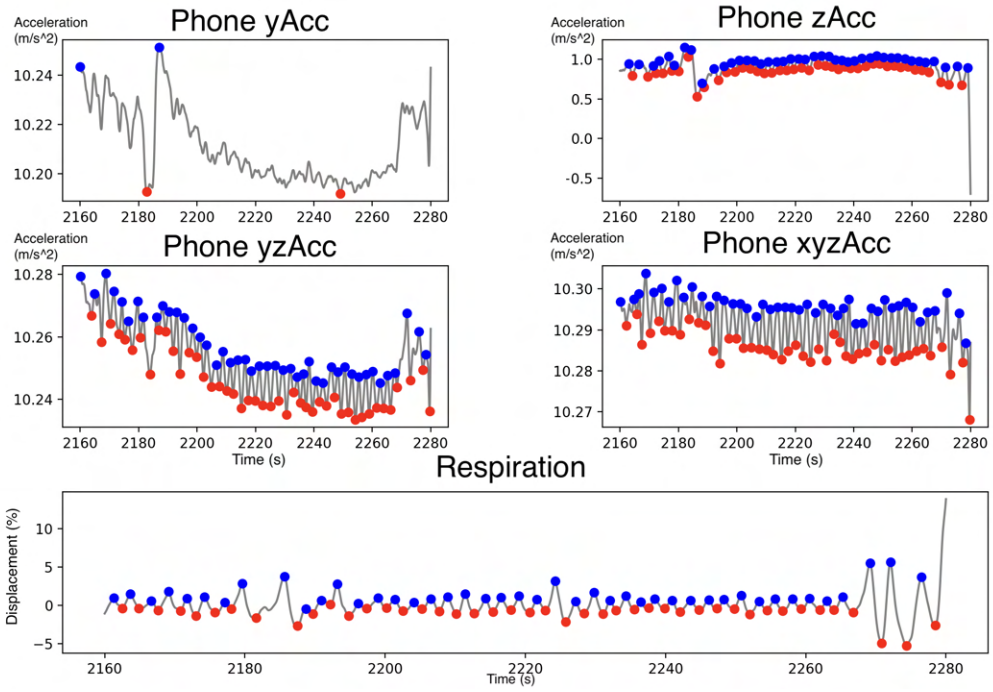


Figure 5.3: Graphical representation of the y-axis, z-axis, yz-combined-axes, xyz-combined-axes and respiration signal provided by the Inductive Respiration sensor on a mild stressful period.

The x-axis was not considered when building the algorithm, since it is unfit for movement detection in desirable directions (it only registers the acceleration horizontally) meaning that this component would never be better than the remaining ones. The y-axis proved not to be ideal as well, but it is more sensitive to respiration movements than the x-axis, as the chest-attached smartphone moves slightly upward and downward. In comparison to the comfortable slow breathing and fast breathing, in the mildly-stressed breathing period, the y-axis poorly captured the breathing pattern. Since the subject is not as still and as focused on breathing when distracted by other activities, the y-component of the ACC is susceptible to more artifacts due to natural motion, providing very different results from those obtained by the RIP sensor. All these factors together infer that the y-axis does not provide a reliable standalone respiration detection. As the z-axis captures the person's forward and backward acceleration, it is the most promising axis for RR detection. Both in comfortable slow and in fast breathing this axis proved to be the best out of the four options estimating the number of peaks and the RR. In the mild stress period, this axis performed much better than the y-axis in peak detection and, consecutively, in estimating the RR. The combination of the y and z axes proved to be the best choice for estimating the number of peaks and the RR in the period that most closely resembles the stress experienced by a worker. This combination also proved to be quite satisfactory in the other two periods. However, since the x-axis also plays a part on enhancing the respiration signal by reducing its overall noise, the combination chosen to estimate the RR was the one that uses all the axes. This combination allowed a close estimation of the number of peaks and RR throughout the three different periods distinguished in the study, even though it was not the top combination in any of them.

Overall it can be concluded that, although the z-axis and the combination of the y and z axes yielded better results, the fusion of all axes was a safer close second choice, since the existence of the x-axis can assist in the eradication of artifacts produced by the other two axes.

5.2 Machine Learning Classification Models

ML models were initially developed using a sample distribution of 75% for training and 25% for testing. In order to verify if the algorithms were performing at a decent level (>80%), the test set was expanded to 40% of the original data set. All metrics discussed in the following sections are related to the results attained for the 40% test data set. The CM for each one of the subsequent best tested models will also be shown, with TP, TN (both colored in light yellow), FP, and FN arranged in the same order as in Section 2.5. Tables with the evaluation metrics, also explained in Section 2.5, are presented (the orange colored cells highlight which classifier obtained the best result within each evaluation metric) as well as the maximum, minimum, mean (colored in light blue) and standard deviation obtained when testing each model with each classifier.

5.2.1 Perceived Stress Scale

Regarding the PSS, an individual's score can range between 0 and 40, with higher values indicating more felt stress [87]. In the first study, 50% of individuals had a score below 13, indicating low stress, and the remaining ones had a score between 14-26, indicating moderate stress as shown in Tables 5.4 and 5.5. In the second study, however, only four people out of twelve reported low stress, while the remaining participants reported moderate stress.

The RF classifier showed that there was some correlation between this score and the person's stress level. The SVM, on the other hand, did not reveal a substantial association, as shown in Tables 4.7 and 4.6 of Section 4.5.1. Because the KNN cannot offer information on features as mentioned in Section 2.5.3, it was unable to determine if the PSS has any influence on the stress detection. As a result, no significant correlation was found between PSS score and the detection of stress levels in participants.

Table 5.4: PSS Score in Stress Study.

Stress Level	Score	Nr. of Participants
Low	0-13	6
Moderate	14-26	6
High	27-40	0

Table 5.5: PSS Score in Calm Study.

Stress Level	Score	Nr. of Participants
Low	0-13	4
Moderate	14-26	8
High	27-40	0

5.2.2 Heart-Rate Model

Due to the smartwatch constraint indicated in Section 4.3.1.1, two techniques were used to increase the quantity of acquired data: re-sampling and overlapping sliding windows for feature extraction. Although this pre-processing enabled information gain, it also had a downside. All features calculated in the frequency domain had to be discarded because this information was lost so that data gain could exist. Fortunately, some studies [3, 103] have concluded that time domain features provide more than enough information to distinguish stressful moments. That was also confirmed with the obtained accuracy. Looking at Table 5.6, one can observe that the KNN algorithm obtained the best mean accuracy result of 87.4% and the lowest standard deviation value of 0.6% in comparison to the SVM and RF, which obtained a mean accuracy of 56.1% and 85.7% with a standard deviation of 2.4% and 2.3%, respectively.

In Table 5.7 one can observe that the SVM classifier got a lot of FP, i.e. uncorrectly instances of "Not Stressed" classified into "Stressed". The most important thing for all developed models is to have as much stress detection as possible, as there is not a serious risk associated with a bad detection of stress (as for example in a false detection of a cancer). It is of greater interest that developed models manage to identify as many cases of stress as possible without a large associated cost when there is a false detection of stress. As a result, it is appropriate to look at the obtained recall value (Table 5.8), which gives

Table 5.6: Accuracy values obtained for the heart rate model.

		Accuracy			
Signal	**Classifiers**	**Max (%)**	**Min (%)**	**Mean (%)**	**STD (%)**
Heart Rate	SVM	59.1%	52.5%	56.1%	2.4%
	RF	88.7%	82.6%	85.7%	2.3%
	KNN	87.8%	86.4%	87.4%	0.6%

the ratio between true positives and predicted negatives, and verifies that SVM algorithm obtained a value of 67.0%. Still looking at recall, the RF had a 88.9% value and the KNN had 92.8%, both for the average of all tests. The RF classifier stood out among the rest regarding the precision evaluation as well as the specificity. The KNN algorithm attained the higher values for the recall, negative predictivity, and accuracy.

Table 5.7: Best confusion matrix for the heart rate model.

			Classification	
Signal	**Classifiers**	**Actual**	**Not Stressed**	**Stressed**
Heart Rate	SVM	Not Stressed	69	96
		Stressed	45	135
	RF	Not Stressed	130	35
		Stressed	4	176
	KNN	Not Stressed	148	17
		Stressed	25	155

Table 5.8: Evaluation metrics for the heart rate model.

		Evaluation Metrics (%)				
Signal	**Classifiers**	**Recall**	**Precision**	**Specificity**	**Negative Predictivity**	**Accuracy**
Heart Rate	SVM	67.0%	56.7%	44.2%	55.4%	56.1%
	RF	88.9%	87.3%	85.7%	87.4%	87.9%
	KNN	92.8%	87.0%	84.8%	91.5%	89.0%

Both RF and KNN were considered strong algorithms for stress classification in the HR model, as opposed to SVM which had the lowest rates in all metrics. The KNN was rated as the best in terms of stress classification since it had the highest evaluation metrics.

5.2.3 Respiration Model

In this model, the mean accuracy of each classifier (shown in Table 5.9) was 81.2% for the SVM, 98.5% for the RF, 98.3% for the KNN and the standard deviations were 2.5%, 0.9% and 0.1%, respectively. The KNN, like in the HR model, has the lowest standard deviation which means it is a very stable algorithm when making predictions.

Table 5.9: Accuracy values obtained for the respiration model.

		Accuracy			
Signal	**Classifiers**	**Max (%)**	**Min (%)**	**Mean (%)**	**STD (%)**
Respiration	SVM	83.0%	76.6%	81.2%	2.5%
	RF	100%	97.7%	98.5%	0.9%
	KNN	99.5%	96.3%	98.3%	0.1%

Table 5.10 shows that SVM attained some FP and FN cases, whereas RF and KNN have zero instances of FN and only one detection of a FN instance in the KNN CM. Furthermore, as can be observed by looking at Table 5.11, SVM is once more distinguishable from the other two algorithms by obtaining the lowest rates in all metrics, even though the attained rate was deemed minimally satisfactory (>75%). It was then considered that, for the respiration signal model, both the RF and the KNN algorithms exhibited promising outcomes. RF performed superiorly regarding the mean accuracy measured across several tests, making it the best classifier for the respiration model's stress classification.

Table 5.10: Best confusion matrix for the respiration model.

			Classification	
Signal	**Classifiers**	**Actual**	**Not Stressed**	**Stressed**
Respiration	SVM	Not Stressed	88	26
		Stressed	11	93
	RF	Not Stressed	114	0
		Stressed	0	104
	KNN	Not Stressed	113	1
		Stressed	0	104

Table 5.11: Evaluation metrics for the respiration model.

		Evaluation Metrics (%)				
Signal	**Classifiers**	**Recall**	**Precision**	**Specificity**	**Negative Predictivity**	**Accuracy**
Respiration	SVM	84.2%	78.5%	78.5%	84.5%	81.2%
	RF	98.1%	98.9%	98.9%	98.3%	98.5%
	KNN	98.3%	98.3%	98.4%	98.4%	98.3%

5.2.4 EMG Model

For both the right and left sides, the accuracy was quite high, as can be seen in Tables 5.12 and 5.13. The left EMG attained a mean accuracy of 85.3%, 98.0% and 98.9%, with standard deviations of 0.0%, 0.0% and 0.8%, respectively and the right EMG got 92.3%, 97.6% and 99.2% for SVM, RF and KNN and the respective standard deviations for each one were 0.0%, 0.0% and 0.0%. Overall, all models proved to be very consistent having a zero standard deviation value for almost all of them. KNN showed best outcomes detecting stress for both the left and the right EMG.

Table 5.12: Accuracy values obtained for the left electromyography model.

		Accuracy			
Signal	**Classifiers**	**Max (%)**	**Min (%)**	**Mean (%)**	**STD (%)**
Left EMG	SVM	88.1%	81.2%	85.3%	0.0%
	RF	100%	96.3%	98.0%	0.0%
	KNN	100%	97.8%	98.9%	0.8%

Table 5.13: Accuracy values obtained for the right electromyography model.

		Accuracy			
Signal	**Classifiers**	**Max (%)**	**Min (%)**	**Mean (%)**	**STD (%)**
Right EMG	SVM	94.5%	90.8%	92.3%	0.0%
	RF	99.1%	95.0%	97.6%	0.0%
	KNN	100%	97.2%	99.2%	0.0%

Additionally, as verified for the HR and respiration models, for both the left and the right EMG, SVM was the algorithm that exhibited the maximum identification of FP and FN (shown in Tables 5.14 and 5.15). Regarding the evaluation metrics revealed in Tables 5.16 and 5.17, the KNN classifier achieved the highest rates for all metrics for both the left EMG, and the right EMG.

Table 5.14: Best confusion matrix for the left electromyography model.

			Classification	
Signal	**Classifiers**	**Actual**	**Not Stressed**	**Stressed**
Left EMG	SVM	Not Stressed	99	15
		Stressed	11	93
	RF	Not Stressed	114	0
		Stressed	0	104
	KNN	Not Stressed	114	0
		Stressed	0	104

Table 5.15: Best confusion matrix for the right electromyography model.

			Classification	
Signal	**Classifiers**	**Actual**	**Not Stressed**	**Stressed**
Right EMG	SVM	Not Stressed	108	6
		Stressed	6	98
	RF	Not Stressed	102	2
		Stressed	0	104
	KNN	Not Stressed	114	0
		Stressed	0	104

Table 5.16: Evaluation metrics for the left electromyography model.

		Evaluation Metrics (%)				
Signal	**Classifiers**	**Recall**	**Precision**	**Specificity**	**Negative Predictivity**	**Accuracy**
Left EMG	SVM	86.5%	83.5%	84.2%	87.4%	85.3%
	RF	99.4%	96.5%	96.7%	99.5%	98.0%
	KNN	99.6%	98.1%	98.2%	99.6%	98.9%

Table 5.17: Evaluation metrics for the right electromyography model.

		Evaluation Metrics (%)				
Signal	**Classifiers**	**Recall**	**Precision**	**Specificity**	**Negative Predictivity**	**Accuracy**
Right EMG	SVM	92.5%	91.5%	92.1%	93.1%	92.3%
	RF	98.1%	97.1%	97.1%	98.3%	97.6%
	KNN	98.3%	100%	100%	98.5%	99.2%

Regarding the combined EMG, Table 5.18 shows that the SVM algorithm obtained a 95.8% mean accuracy with a 1.2% standard deviation, RF attained a 98.6% with a 0.9% standard deviation and KNN achieved again the higher mean accuracy value of 9934% with a 0.6% standard deviation.

Table 5.18: Accuracy values obtained for the total electromyography model.

		Accuracy			
Signal	**Classifiers**	**Max (%)**	**Min (%)**	**Mean (%)**	**STD (%)**
Total EMG	SVM	96.8%	94.5%	95.8%	1.2%
	RF	100%	97.7%	98.6%	0.9%
	KNN	99.5%	98.2%	99.3%	0.6%

In its CM, the RF classifier obtained 0 FP and FN during one of the tests, the KNN detected 1 incidences of FN and the SVM identified 4 cases of FN and 4 of FP (presented in Table 5.19). Table 5.20 demonstrates that RF and KNN perform quite well in this model, but KNN stands out. Regarding the mean accuracies obtained for each algorithm,

KNN was also the one with a higher rate.

Table 5.19: Best confusion matrix for the total electromyography model.

			Classification	
Signal	**Classifiers**	**Actual**	**Not Stressed**	**Stressed**
Total EMG	SVM	Not Stressed	110	4
		Stressed	4	100
	RF	Not Stressed	114	0
		Stressed	0	104
	KNN	Not Stressed	114	0
		Stressed	1	103

Table 5.20: Evaluation metrics for the total electromyography model.

		Evaluation Metrics (%)				
Signal	**Classifiers**	**Recall**	**Precision**	**Specificity**	**Negative Predictivity**	**Accuracy**
Total EMG	SVM	94.4%	96.7%	97.0%	95.0%	95.8%
	RF	99.4%	97.7%	99.9%	99.5%	98.6%
	KNN	98.5%	100%	100%	98.6%	99.3%

For the total EMG and looking at the features in bold in Table 4.6 of Section 4.5.1, it's observable that features from the right EMG were better for all classifiers. Out of the 10 best ranked features, the ones that come from the right EMG are in bold and in both the RF and the SVM classifiers 8 out of 10 features were selected from the right EMG. One might conclude that the accuracy of the right EMG is substantially higher than that of the left one by comparing the accuracy values yielded by the SVM, RF, and KNN for the right EMG signal to those found on the same classifiers for the left EMG signal. All participants were right-handed and they used the computer's trackpad far more frequently during the stress protocol than during the relaxation protocol which can be a very likely explanation for this accuracy discrepancy. In other words, the stress protocol contains significantly more muscle events than the relaxing protocol, and the stress assessment may be based on this.

Overall, it was found that KNN was the best algorithm to classify stress for the left EMG, the right EMG and the total EMG.

5.2.5 Multimodal Model

To make the multimodal model more reliable, only features from the left EMG were used, because information coming from that signal was less prone to be adjusted to the protocols than information coming from the right EMG.

The results for this model were really promising for all algorithms. Regarding the mean accuracies obtained after testing the classifiers 5 times (Table 5.21), the SVM performed with a mean accuracy of 96,4% and a standard deviation of 1.0%; the RF obtained a mean accuracy of 99,4% and a standard deviation of 0.5% and finally, the KNN revealed a mean accuracy of 100% and a standard deviation of 0.0%, being once again the best algorithm and obtaining a perfect score in stress classification.

Table 5.21: Accuracy values obtained for the multimodal model.

		Accuracy			
Signal	**Classifiers**	**Max (%)**	**Min (%)**	**Mean (%)**	**STD (%)**
Multimodal	SVM	97.1%	95.1%	96.4%	1.0%
	RF	100%	98.8%	99.4%	0.5%
	KNN	100%	100%	100%	0%

Table 5.22 reveals that both KNN and RF algorithms correctly detected all instances of "Stressed" and "Not Stressed" and the SVM incorrectly identified 4 FN and 4 FP. In Table 5.23 one can see that both the KNN and the RF reached the highest achievable rates for all analyzed metrics.

Table 5.22: Best confusion matrix for the multimodal model.

			Classification	
Signal	**Classifiers**	**Actual**	**Not Stressed**	**Stressed**
Multimodal	SVM	Not Stressed	161	4
		Stressed	4	176
	RF	Not Stressed	165	0
		Stressed	0	180
	KNN	Not Stressed	165	0
		Stressed	0	180

Table 5.23: Evaluation metrics for the multimodal model.

		Evaluation Metrics (%)				
Signal	**Classifiers**	**Recall**	**Precision**	**Specificity**	**Negative Predictivity**	**Accuracy**
Multimodal	SVM	96.30%	97.1%	96.8%	95.7%	96.4%
	RF	99.6%	99.3%	99.3%	99.5%	99.4%
	KNN	100%	100%	100%	100%	100%

Even though this model's result were extraordinarily high, it stands to reason that a multimodal model that incorporates the best features of three separate physiological signals may recognize stressful moments with such certainty. KNN was the most successful classifier for stress detection, displaying 100% accuracy throughout all tests.

5.3 Comparison Between Models

5.3.1 Best Models

As stated in the aforementioned sections, the SVM algorithm, despite most of the time giving a satisfactory stress rating (>80%), proved to consistently have the lowest performances. Since the other two techniques are non-linear and the SVM algorithm uses a linear kernel, it can be deduced that adopting linearity for stress classification is not the best option for the created models. The performance of KNN and RF was fairly identical, with KNN slightly standing out as can be seen in the graphic of Figure 5.4. The KNN approach, which classifies an instance based on neighbors with similar properties, or the RF strategy, which combines the classifications generated by various decision trees and reaches a majority decision, are techniques that were well suited in classifying stress for the selected biosignals with the chosen protocols. Unfortunately, since participants were given clear instructions on how to wear all the equipment and were aware of its existence during the protocols, the obtained accuracies would indubitably decrease in real-time detection. There are no limitations on movement in daily life and people usually perform many tasks at once, which makes detection more complex. Additionally, recording these physiological signals in an uncontrolled environment can be challenging due to a variety of factors, that affect physiology, other than stress. A real-time stress detection performance could deteriorate as a result of these two challenges. In order to circumvent them, one possible solution would be to add more pre-processing steps to the acquired signals and attempt to retrieve them in conditions as similar as possible to those experienced by the worker in order to analyze more precisely the effects of stress on computer users.

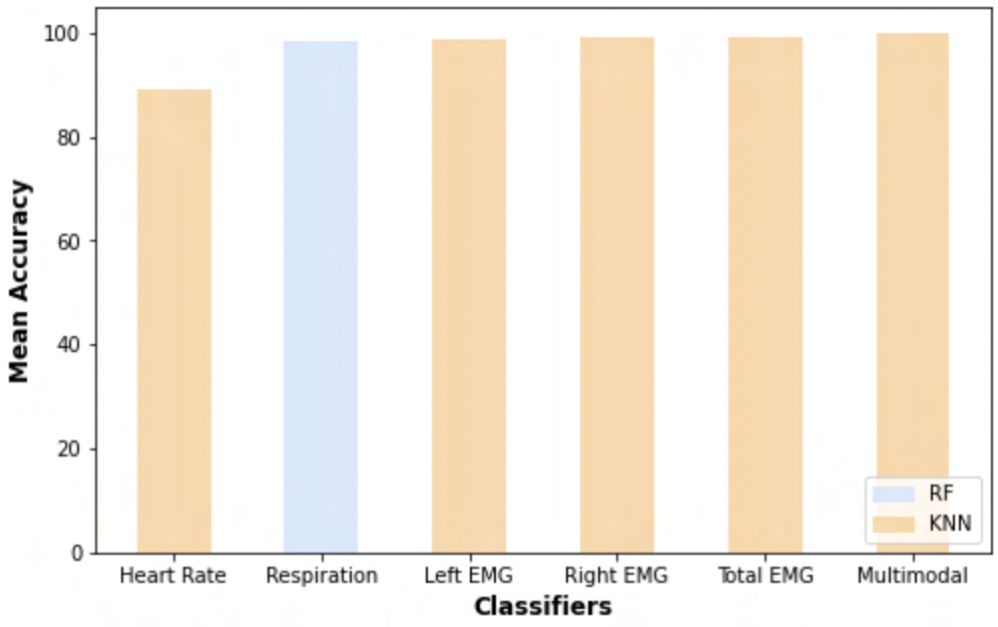


Figure 5.4: Best algorithm for each model.

Regarding each specific model, the accuracies of the three different types of classifiers were lower for the HR signal in comparison to the respiration and EMG. This model was the one with the poorest results for every classifier and several reasons can be behind this. The most obvious one is that this model, even with the data-increasing techniques, had access to a much smaller data set compared to the others. In addition to being the model that was developed with less data, it was also the one that used the fewest features. This model was only trained with the 4 features shown in blue in Table 4.7 from Section 4.5.1 as they were the best. Respiration was the second model with less information after the HR model, but the difference between the two is still significant (respiration has twice the HR data) justifying why metrics were comparatively better. As the EMG was the signal with the most data and the most features, it was expected that the algorithms would obtain satisfactory accuracies, and this effectively happened. Moreover, the multimodal model only used data from the periods in which all sensors acquired at the same time (Section 4.4.2.4). Thus, the amount of information used to develop the multimodal model was significantly less than the amount used, for the left, right, and total EMG, the three of which did not obtain results as high as the multimodal model. It can be speculated that, even with less data, a model that uses multiple modalities to recognize a stressful situation may perform better than unimodal models because it accesses different physiological responses produced by the stressed individual.

5.3.2 Related Work

Referring to prior studies and comparing the accuracy of various stress detection approaches, the selected biosignals, and processing techniques were found to be very efficient attaining higher accuracies (96.4%, 99.4%, and 100% for the SVM, RF and KNN, respectively) compared to those obtained in almost all of the studies mentioned in the Table 5.24.

In this work, the models that were created are generic, meaning that they can be used to analyze data from any individual. Some research [60, 61, 69, 74] built customized models for each person, consistently outperforming generalized models in assessments. Considering this, the accuracy rates obtained with the created generalized models are highly promising. The personalized approach, however, has the advantage of discovering precisely how each individual's physiological signals function, making it always more tailored than the generalized approach. Furthermore, both cognitive and emotional stress were examined, which was something that related studies lacked, and which is significant because these two types of stress are the most common in computer users [25].

Table 5.24: Comparison of obtained accuracy with ones obtained by other studies

Study	Biosignals	Classifier	Accuracy
Wijsman et al. [67]	HR, Respiration, EMG, GSR	GEE	74.5%
Wijsman et al. [65]	ECG, Respiration, EMG, SC	LBN, QBN, KNN, FLSL	≈80.0%
Zhai et al. [10]	BVP, GSR, PD	SVM	80.0%
Choi et al. [25]	HR, EMG, EDA	LR	81.0%
Smets et al. [71]	EMG, GSR, ST	SVM	82.7%
Ciabattoni et al. [2]	HR, GSR, ST	KNN	84.5%
Siirtola et al. [11]	HR, BVP, ST	SVM	87.4%
Zhai et al. [68]	BVP, GSR, PD, ST	SVM	90.1%
Majid et al. [63]	PPG, EEG, GSR,	MPL	95.0%
Pourmohammadi et al. [17]	ECG, EMG	SVM	100%
This thesis	HR, Respiration, EMG	SVM RF KNN	96.4% 99.4% 100%

6

Conclusion

6.1 General Results

The Laboratory for Instrumentation, Biomedical Engineering, and Radiation Physics (LIBPhys) at FCT-UNL was where this master's thesis was developed. The primary goal of this project was to create stress detection classifiers using ML algorithms to analyze WRS in computer users. Therefore, several steps were done in order to develop this work. The first was to create an algorithm as robust as possible that could accurately estimate the RR of each individual while working, using an ACC sensor from a smartphone placed on an individual's chest. This was successfully achieved, as shown in Section 5.1. Then, two studies were conducted to collect the data that would later be employed by the classification systems. The protocols were successfully carried out allowing for a vast supply of biosignals: HR, provided by the PPG smartwatch sensor; respiration, derived from a smartphone ACC placed on the chest; and trapezius EMG, using two muscleBANs. It was important to look at both cognitive and emotional stress because computer users tend to experience these two types of stress most frequently while working. Unfortunately, the information obtained from the HCI with the *Latent* [16] tool was insufficient to draw any conclusions, and it was not possible to extract further data using the *Latent* tool, due to installation complications. With the acquired data, features were firstly extracted and secondly selected in order to train three individual models, one for each physiological signal and a fused multimodal model, combining the best features of the others. Using SVM, RF, and KNN as classification algorithms, the models were assessed with different evaluation metrics. The HR model reached accuracy rates of 56.1%, 85.7% and 87.4% for the SVM, RF and KNN, respectively; and was the model that was less successful in the classification process. The respiration model got accuracy rates of 81.2% for the SVM, 98.5% for the RF, 98.3% for the KNN. Regarding the EMG model accuracy rates for each classifier also in the aforementioned order were 95.8%, 98.6% and 99.3% combining features from both left and right trapezius. Finally, the multimodal model had accuracies of 96.4% with the SVM algorithm, 99.4% with RF and 100% KNN. This model used the best features extracted from the HR, the respiration and the left EMG biosignals.

When compared to other studies that employed multimodal approaches for mental stress classification, the findings in this work were considerably better. However, the obtained accuracies would certainly be reduced in a real-time detection because throughout the protocols, participants were clearly instructed on how to properly wear the equipment and they were aware of its existence. In daily life, there are no restrictions on movement. People frequently engage in many activities at once, which complicates the detection process. Motion artifacts could rise and stress detection systems' performance might suffer as a result.

6.2 Future Work

As previously mentioned, due to *Latent*'s [16] installation limitations, an evaluation of the stress effect during HCI was not pursued, but it would be an interesting future approach to explore.

Regarding the created respiration algorithm, although it can successfully identify peaks and valleys caused by the expansion and contraction of the diaphragm, its sensitivity to movement can still be improved.

Also, in this work, an early fusion of signals was made in order to create the multimodal model. Alternatively, another possible approach could be to train a classifier for each biosignal and conduct a majority vote utilizing the output from all classifiers.

Moreover, this stress detection algorithm only functioned in offline mode, which means it collected data in real-time but processed the data offline. However, online detection of stressful periods can enable the delivery of appropriate interventions at the right time when the individual is experiencing physiological stress. An online model, like the one in Figure 6.1, would be an approach that would be interesting to explore in the future. Thus, techniques for detecting stress in real-time are required. Whenever stressful moments are observed, real-time detection will allow researchers to implement just-in-time adaptive interventions (JITAIs).

Monitoring stress in real-time conditions can provide direct biofeedback to the user and allow for early self-intervention [9]. This can enhance an individual's self-awareness. There is a user tendency to adopt coping strategies, not only by using the interventions suggested but also by understanding that simple activities can actually help them manage their stress [107].

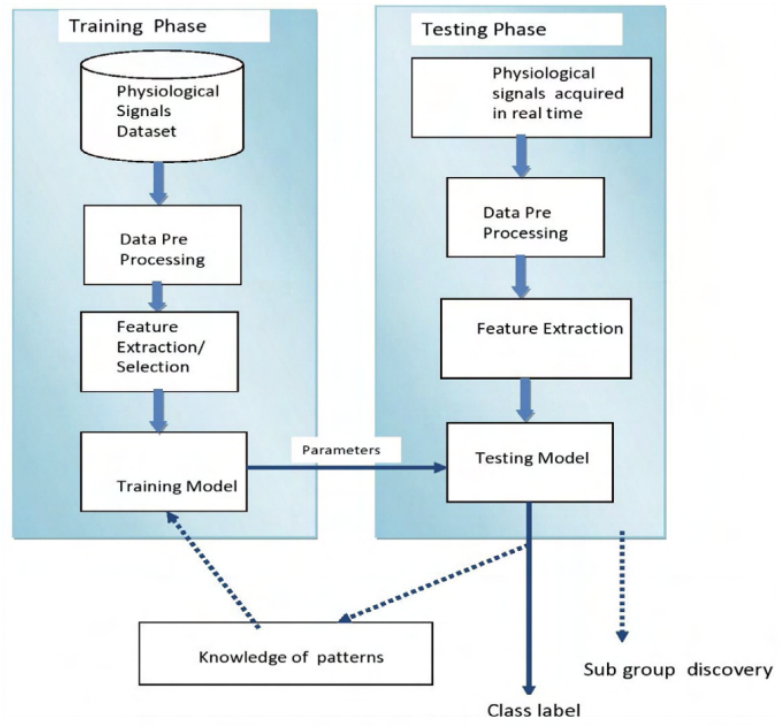


Figure 6.1: Schematic of proposed future model. Retrieved from [76].

6.3 Ethical Issues

Measuring computer workers stress levels can be beneficial to both users and companies. As discussed in the above Section, a future prospect of this thesis would be to implement the developed models in real-time working scenarios.

However, data management must be done with great caution. The models created for this thesis help identify sensitive information about the person. If they are to be used in a real context, the user must be aware of the advantages of collecting specific data about their physiological signals for a given purpose and asked for permission to process his/her biosignals. Additionally, for security reasons, the computer user should always have the ability to erase the data. Keeping personal data without specific consent is illegal, according to the General Data Protection Regulation (GDPR) [108]. Knowing which specific individual the data belongs to can potentially lead to improper use of this information. To overcome this ethical issue, one must also ensure that personal data cannot be traced back. To maintain personal data anonymous, data can be pooled in statistical metrics such as mean, median, percentiles and others.

Bibliography

[1] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/master/template.pdf (cit. on p. iii).

[2] L. Ciabattoni et al. "Real-time mental stress detection based on smartwatch". In: Institute of Electrical and Electronics Engineers Inc., 2017-03, pp. 110–111. ISBN: 9781509055449. DOI: 10.1109/ICCE.2017.7889247 (cit. on pp. 1, 21, 73).

[3] M. Salai, I. Vassányi, and I. Kósa. "Stress detection using low cost heart rate sensors". In: *Journal of healthcare engineering* 2016 (2016) (cit. on pp. 1, 18–20, 46, 64).

[4] S. Saxena, P. Saxena, and S. K. Dubey. "Various levels of human stress and their impact on human computer interaction". In: *2013 International Conference on Human Computer Interactions (ICHCI)*. IEEE Computer Society, 2013, pp. 1–6. ISBN: 9781467357036. DOI: 10.1109/ICHCI-IEEE.2013.6887808 (cit. on pp. 1, 24).

[5] K. Daher et al. "Reduce Stress Through Empathic Machine to Improve HCI". In: vol. 1152 AISC. Springer, 2020, pp. 232–237. ISBN: 9783030442668. DOI: 10.1007/978-3-030-44267-5_35 (cit. on pp. 1, 2, 24, 25).

[6] Y. S. Can, B. Arnrich, and C. Ersoy. "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey". In: vol. 92. Academic Press Inc., 2019-04. DOI: 10.1016/j.jbi.2019.103139 (cit. on pp. 1–3, 6, 10, 18, 33, 34, 37).

[7] S. Clarke, L. G. Jaimes, and M. A. Labrador. "mstress: A mobile recommender system for just-in-time interventions for stress". In: *2017 14th IEEE Annual Consumer Communications and Networking Conference (CCNC)*. IEEE. 2017, pp. 1–5 (cit. on p. 1).

[8] F. Goncalves et al. "X3S: A multi-modal approach to monitor and assess stress through human-computer interaction". In: *Computer Science and Information Systems* 15 (3 2018-10), pp. 683–703. ISSN: 24061018. DOI: 10.2298/CSIS180115033G (cit. on pp. 1, 24).

[9] T. Chalmers et al. "Stress watch: The use of heart rate and heart rate variability to detect stress: A pilot study using smart watch wearables". In: *Sensors* 22 (1 2022-01). ISSN: 14248220. DOI: 10.3390/s22010151 (cit. on pp. 1, 3, 76).

[10] J. Zhai et al. "Realization of stress detection using psychophysiological signals for improvement of human-computer interactions". In: *Proceedings. IEEE SoutheastCon, 2005.* IEEE. 2005, pp. 415–420 (cit. on pp. 1, 2, 21, 73).

[11] P. Siirtola. "Continuous stress detection using the sensors of commercial smartwatch". In: Association for Computing Machinery, Inc, 2019-09, pp. 1198–1201. ISBN: 9781450368698. DOI: 10.1145/3341162.3344831 (cit. on pp. 1, 2, 21, 73).

[12] J. Takala and M. Iscar Urrutia. "Safety and health at work: a European perspective". In: *Revista Portuguesa de Saúde Pública* (2009), pp. 21–30 (cit. on pp. 1, 2).

[13] M. Kivimäki and I. Kawachi. "Work stress as a risk factor for cardiovascular disease". In: *Current cardiology reports* 17.9 (2015), pp. 1–9 (cit. on p. 2).

[14] S. Baltaci and D. Gokcay. "Stress Detection in Human–Computer Interaction: Fusion of Pupil Dilation and Facial Temperature Features". In: *International Journal of Human-Computer Interaction* 32 (12 2016-12), pp. 956–966. ISSN: 15327590. DOI: 10.1080/10447318.2016.1220069 (cit. on p. 2).

[15] M. J. Smith, B.-T. Karsh, and F. B. Moro. "A review of research on interventions to control musculoskeletal disorders". In: *Work-Related Musculoskeletal Disorders: Report, Workshop Summary, and Workshop Papers, National Research Council, National Academy of Sciences*. National Academy Press Washington, DC. 1999, pp. 200–229 (cit. on p. 2).

[16] C. Cepeda et al. "Latent: A Flexible Data Collection Tool to Research Human Behavior in the Context of Web Navigation". In: *IEEE Access* 7 (2019), pp. 77659–77673. ISSN: 21693536. DOI: 10.1109/ACCESS.2019.2916996 (cit. on pp. 2, 3, 11, 26, 27, 35, 43, 54, 75, 76).

[17] S. Pourmohammadi and A. Maleki. "Stress detection using ECG and EMG signals: A comprehensive study". In: *Computer methods and programs in biomedicine* 193 (2020), p. 105482 (cit. on pp. 3, 10, 18, 22, 33, 38, 51, 73).

[18] E. Hehman, R. M. Stolier, and J. B. Freeman. "Advanced mouse-tracking analytic techniques for enhancing psychological science". In: *Group Processes and Intergroup Relations* 18 (3 2015-05), pp. 384–401. ISSN: 14617188. DOI: 10.1177/1368430214538325 (cit. on p. 3).

[19] S. O. Rajankar and S. N. Talbar. "An electrocardiogram signal compression techniques: a comprehensive review". In: *Analog Integrated Circuits and Signal Processing* 98.1 (2019), pp. 59–74 (cit. on pp. 5, 6).

[20] K. Ranjeet, A. Kumar, and R. K. Pandey. "ECG signal compression using different techniques". In: *International Conference on Advances in Computing, Communication and Control*. Springer. 2011, pp. 231–241 (cit. on p. 5).

[21] J. Parák and J. Havlík. "ECG signal processing and heart rate frequency detection methods". In: *Proceedings of Technical Computing Prague* 8 (2011), p. 2011 (cit. on pp. 5, 6).

[22] L. Sörnmo and P. Laguna. "Electrocardiogram (ECG) signal processing". In: *Wiley encyclopedia of biomedical engineering* (2006) (cit. on p. 6).

[23] P. S. Pandey. "Machine learning and IoT for prediction and detection of stress". In: *2017 17th International Conference on Computational Science and Its Applications (ICCSA)*. IEEE. 2017, pp. 1–5 (cit. on p. 6).

[24] U. Satija, B. Ramkumar, and M. S. Manikandan. "A review of signal processing techniques for electrocardiogram signal quality assessment". In: *IEEE reviews in biomedical engineering* 11 (2018), pp. 36–52 (cit. on pp. 6, 13).

[25] J. Choi, B. Ahmed, and R. Gutierrez-Osuna. "Development and evaluation of an ambulatory stress monitor based on wearable sensors". In: *IEEE transactions on information technology in biomedicine* 16.2 (2011), pp. 279–286 (cit. on pp. 6–8, 18, 23, 36, 72, 73).

[26] T. Bhowmik, J. Dey, and V. N. Tiwari. "A novel method for accurate estimation of HRV from smartwatch PPG signals". In: *2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE. 2017, pp. 109–112 (cit. on pp. 6, 7, 34).

[27] L. M. Nilsson. "Respiration signals from photoplethysmography". In: *Anesthesia & Analgesia* 117.4 (2013), pp. 859–865 (cit. on p. 7).

[28] A. Pinge et al. "A Comparative Study between ECG-based and PPG-based Heart Rate Monitors for Stress Detection". In: *2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS)*. IEEE. 2022, pp. 84–89 (cit. on pp. 7, 33, 46).

[29] N. Pinheiro et al. "Can PPG be used for HRV analysis?" In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2016, pp. 2945–2949 (cit. on pp. 7, 46).

[30] A. Pedrana et al. "Development of a wearable in-ear PPG system for continuous monitoring". In: *IEEE Sensors Journal* 20.23 (2020), pp. 14482–14490 (cit. on p. 7).

[31] S. Preejith et al. "Accelerometer based system for continuous respiratory rate monitoring". In: *2017 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE. 2017, pp. 171–176 (cit. on pp. 8, 9, 31, 59).

[32] R. A. Hameed et al. "Human emotion classification based on respiration signal". In: *Proceedings of the International Conference on Information and Communication Technology*. 2019, pp. 239–245 (cit. on p. 8).

[33] A. Bates et al. "Respiratory rate and flow waveform estimation from tri-axial accelerometer data". In: *2010 International Conference on Body Sensor Networks*. IEEE. 2010, pp. 144–150 (cit. on p. 9).

[34] L. Estrada et al. "Respiratory signal derived from the smartphone built-in accelerometer during a Respiratory Load Protocol". In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2015, pp. 6768–6771 (cit. on p. 9).

[35] P. D. Hung. "Estimating respiration rate using an accelerometer sensor". In: *Proceedings of the 8th International Conference on Computational Systems-Biology and Bioinformatics*. 2017, pp. 11–14 (cit. on pp. 9, 59).

[36] M. B. I. Reaz, M. S. Hussain, and F. Mohd-Yasin. "Techniques of EMG signal analysis: detection, processing, classification and applications". In: *Biological procedures online* 8.1 (2006), pp. 11–35 (cit. on pp. 9, 10).

[37] W. Chen et al. "Characterization of surface EMG signal based on fuzzy entropy". In: *IEEE Transactions on neural systems and rehabilitation engineering* 15.2 (2007), pp. 266–272 (cit. on p. 10).

[38] P. K. Artemiadis and K. J. Kyriakopoulos. "Teleoperation of a robot manipulator using EMG signals and a position tracker". In: *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2005, pp. 1003–1008 (cit. on p. 10).

[39] G. Sinha, R. Shahi, and M. Shankar. "Human Computer Interaction". In: 2010, pp. 1–4. ISBN: 9780769542461. DOI: 10.1109/ICETET.2010.85 (cit. on p. 11).

[40] P. Booth. *An introduction to human-computer interaction (psychology revivals)*. Psychology Press, 2014 (cit. on p. 11).

[41] A. Dix. "Human-Computer Interaction". In: *Encyclopedia of Database Systems*. Ed. by L. LIU and M. T. ÖZSU. Boston, MA: Springer US, 2009, pp. 1327–1331. ISBN: 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9_192. URL: https://doi.org/10.1007/978-0-387-39940-9_192 (cit. on p. 11).

[42] J. Tao. *Emotion recognition for human-computer interaction*. 2021, p. 1 (cit. on p. 11).

[43] T. Yamauchi and K. Xiao. "Reading Emotion From Mouse Cursor Motions: Affective Computing Approach". In: *Cognitive Science* 42 (3 2018-04), pp. 771–819. ISSN: 15516709. DOI: 10.1111/cogs.12557 (cit. on p. 11).

[44] R. L. Rosa et al. "A knowledge-based recommendation system that includes sentiment analysis and deep learning". In: *IEEE Transactions on Industrial Informatics* 15.4 (2018), pp. 2124–2135 (cit. on p. 11).

[45] A. Subasi. *Practical Machine Learning for Data Analysis Using Python*. Academic Press, 2020 (cit. on pp. 11, 12).

[46] S. Badillo et al. "An introduction to machine learning". In: *Clinical pharmacology & therapeutics* 107.4 (2020), pp. 871–885 (cit. on pp. 11, 14–17).

[47] Z. Ghahramani. "Unsupervised learning". In: *Summer school on machine learning*. Springer. 2003, pp. 72–112 (cit. on pp. 11, 12).

[48] K. Das and R. N. Behera. "A survey on machine learning: concept, algorithms and applications". In: *International Journal of Innovative Research in Computer and Communication Engineering* 5.2 (2017), pp. 1301–1309 (cit. on p. 12).

[49] K. El Bouchefry and R. S. de Souza. "Learning in big data: Introduction to machine learning". In: *Knowledge discovery in big data from astronomy and earth observation*. Elsevier, 2020, pp. 225–249 (cit. on p. 12).

[50] M. Hossin and M. N. Sulaiman. "A review on evaluation metrics for data classification evaluations". In: *International journal of data mining & knowledge management process* 5.2 (2015), p. 1 (cit. on p. 13).

[51] H. Dalianis. "Evaluation metrics and evaluation". In: *Clinical text mining*. Springer, 2018, pp. 45–53 (cit. on p. 13).

[52] T. Hastie, R. Tibshirani, and J. Friedman. "Additive models, trees, and related methods". In: *The Elements of Statistical Learning*. Springer, 2009, pp. 295–336 (cit. on p. 14).

[53] W. S. Noble. "What is a support vector machine?" In: *Nature biotechnology* 24.12 (2006), pp. 1565–1567 (cit. on p. 14).

[54] Y. Liu, Y. Wang, and J. Zhang. "New machine learning algorithm: Random forest". In: *International Conference on Information Computing and Applications*. Springer. 2012, pp. 246–252 (cit. on p. 15).

[55] NIST. *Euclidean distance*. 2004. URL: https://xlinux.nist.gov/dads/HTML/euclidndstnc.html (visited on 2022-10-10) (cit. on p. 16).

[56] NIST. *Manhattan distance*. 2019. URL: https://xlinux.nist.gov/dads/HTML/manhattanDistance.html (visited on 2022-10-10) (cit. on p. 16).

[57] A. Alberdi, A. Aztiria, and A. Basarab. "Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review". In: *Journal of biomedical informatics* 59 (2016), pp. 49–75 (cit. on pp. 17, 18, 21).

[58] P. Kalra and V. Sharma. "Mental stress assessment using PPG signal a deep neural network approach". In: *IETE Journal of Research* (2020), pp. 1–7 (cit. on pp. 18, 21, 46).

[59] P. Karthikeyan, M. Murugappan, and S. Yaacob. "Analysis of stroop color word test-based human stress detection using electrocardiography and heart rate variability signals". In: *Arabian Journal for Science and Engineering* 39.3 (2014), pp. 1835–1847 (cit. on p. 20).

[60] W. Lawanont et al. "Daily stress recognition system using activity tracker and smartphone based on physical activity and heart rate data". In: *International conference on intelligent decision technologies*. Springer. 2018, pp. 11–21 (cit. on pp. 20, 72).

[61] A. Muaremi, B. Arnrich, and G. Tröster. "Towards measuring stress with smartphones and wearable devices during workday and sleep". In: *BioNanoScience* 3.2 (2013), pp. 172–183 (cit. on pp. 20, 72).

[62] M. Zubair and C. Yoon. "Multilevel mental stress detection using ultra-short pulse rate variability series". In: *Biomedical Signal Processing and Control* 57 (2020), p. 101736 (cit. on pp. 21, 22).

[63] M. Majid, A. Arsalan, and S. M. Anwar. "A Multimodal Perceived Stress Classification Framework using Wearable Physiological Sensors". In: *arXiv preprint arXiv:2206.10846* (2022) (cit. on pp. 22, 33, 73).

[64] P. Karthikeyan, M. Murugappan, and S. Yaacob. "EMG signal based human stress level classification using wavelet packet transform". In: *International Conference on Intelligent Robotics, Automation, and Manufacturing*. Springer. 2012, pp. 236–243 (cit. on p. 22).

[65] J. Wijsman et al. "Towards mental stress detection using wearable physiological sensors". In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2011, pp. 1798–1801 (cit. on pp. 22, 33, 36, 73).

[66] J. Wijsman et al. "Trapezius muscle EMG as predictor of mental stress". In: *ACM transactions on embedded computing systems (TECS)* 12.4 (2013), pp. 1–20 (cit. on pp. 22, 36).

[67] J. Wijsman et al. "Wearable physiological sensors reflect mental stress state in office-like situations". In: *2013 humaine association conference on affective computing and intelligent interaction*. IEEE. 2013, pp. 600–605 (cit. on pp. 22, 36, 73).

[68] J. Zhai and A. Barreto. "Stress detection in computer users based on digital signal processing of noninvasive physiological variables". In: *2006 international conference of the IEEE engineering in medicine and biology society*. IEEE. 2006, pp. 1355–1358 (cit. on pp. 22, 73).

[69] Y. Shi et al. "Personalized stress detection from physiological measurements". In: *International symposium on quality of life technology*. 2010, pp. 28–29 (cit. on pp. 23, 72).

[70] K. Palanisamy, M. Murugappan, and S. Yaacob. "Multiple physiological signal-based human stress identification using non-linear classifiers". In: *Elektronika ir elektrotechnika* 19.7 (2013), pp. 80–85 (cit. on p. 23).

[71] E. Smets et al. "Comparison of machine learning techniques for psychophysiological stress detection". In: *International Symposium on Pervasive Computing Paradigms for Mental Health*. Springer. 2015, pp. 13–22 (cit. on pp. 23, 73).

[72] S. Salmeron-Majadas, O. C. Santos, and J. G. Boticario. "An evaluation of mouse and keyboard interaction indicators towards non-intrusive and low cost affective modeling in an educational context". In: *Procedia Computer Science* 35 (2014), pp. 691–700 (cit. on p. 24).

[73] D. Sun, P. Paredes, and J. Canny. "MouStress: Detecting stress from mouse motion". In: Association for Computing Machinery, 2014, pp. 61–70. ISBN: 9781450324731. DOI: 10.1145/2556288.2557243 (cit. on p. 24).

[74] A. O. Akmandor and N. K. Jha. "Keep the stress away with SoDA: Stress detection and alleviation system". In: *IEEE Transactions on Multi-Scale Computing Systems* 3.4 (2017), pp. 269–282 (cit. on pp. 25, 26, 33, 72).

[75] J. Aigrain et al. "Multimodal stress detection from multiple assessments". In: *IEEE Transactions on Affective Computing* 9.4 (2016), pp. 491–506 (cit. on p. 25).

[76] S. S. Panicker and P. Gayathri. "A survey of machine learning techniques in physiology based mental stress detection systems". In: *Biocybernetics and Biomedical Engineering* 39.2 (2019), pp. 444–469 (cit. on pp. 26, 77).

[77] NumFOCUS. *Pandas*. 2022. URL: https://pandas.pydata.org/ (visited on 2022-10-13) (cit. on p. 27).

[78] NumPy. *NumPy*. 2022. URL: https://numpy.org/ (visited on 2022-10-13) (cit. on p. 27).

[79] SciPy. *SciPy*. 2022. URL: https://scipy.org/ (visited on 2022-10-13) (cit. on p. 27).

[80] M. Barandas et al. "TSFEL: Time series feature extraction library". In: *SoftwareX* 11 (2020), p. 100456 (cit. on pp. 27, 49, 55).

[81] R. Varandas et al. "BIOSIGNALS NOTEBOOKS: AN INNOVATIVE APPROACH FOR BIOSIGNALS EXPERIENCE SHARING". In: *INTED2020 Proceedings*. IATED. 2020, pp. 8194–8200 (cit. on pp. 27, 30, 42).

[82] F. Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830 (cit. on pp. 27, 54, 55, 57).

[83] S. Silva et al. *Assessing Occupational Health with a Cross-Platform Application based on Self-Reports and Biosignals* (cit. on p. 28).

[84] *MuscleBANBE Data Sheet*. PLUX– biosignalsplux. 2015 (cit. on pp. 29, 30).

[85] C. Bogler et al. "Default network activity is associated with better performance in a vigilance task". In: *Frontiers in human neuroscience* 11 (2017), p. 623 (cit. on p. 31).

[86] B. Ahmed et al. "ReBreathe: A calibration protocol that improves stress/relax classification by relabeling deep breathing relaxation exercises". In: *IEEE Transactions on Affective Computing* 7.2 (2015), pp. 150–161 (cit. on p. 33).

[87] S. Cohen, T. Kamarck, R. Mermelstein, et al. "Perceived stress scale". In: *Measuring stress: A guide for health and social scientists* 10.2 (1994), pp. 1–2 (cit. on pp. 33, 64).

[88] A. Sano and R. W. Picard. "Stress recognition using wearable sensors and mobile phones". In: *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE. 2013, pp. 671–676 (cit. on p. 33).

[89] H. J. Hermens et al. "Development of recommendations for SEMG sensors and sensor placement procedures". In: *Journal of electromyography and Kinesiology* 10.5 (2000), pp. 361–374 (cit. on p. 34).

[90] M. Foundation. *XPath*. 2022. URL: https://developer.mozilla.org/en-US/docs/Web/XPath (visited on 2022-10-13) (cit. on p. 36).

[91] M. Loeb, D. H. Holding, and M. A. Baker. "Noise stress and circadian arousal in self-paced computation". In: *Motivation and Emotion* 6.1 (1982), pp. 43–48 (cit. on p. 36).

[92] A.-M. Brouwer and M. A. Hogervorst. "A new paradigm to induce mental stress: the Sing-a-Song Stress Test (SSST)". In: *Frontiers in neuroscience* 8 (2014), p. 224 (cit. on pp. 36, 37).

[93] S. O. Ferreira. "Emotional activation in human beings: procedures for experimental stress induction". In: *Psicologia USP* 30 (2019) (cit. on p. 37).

[94] E. Um et al. "Emotional design in multimedia learning." In: *Journal of educational psychology* 104.2 (2012), p. 485 (cit. on pp. 38, 39).

[95] S. Kurt and K. K. Osueke. "The effects of color on the moods of college students". In: *sage Open* 4.1 (2014), p. 2158244014525423 (cit. on pp. 38, 39).

[96] A. Ghandeharioun and R. Picard. "BrightBeat: Effortlessly influencing breathing for cultivating calmness and focus". In: *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. 2017, pp. 1624–1631 (cit. on p. 38).

[97] R. Jerath et al. "Self-regulation of breathing as a primary treatment for anxiety". In: *Applied psychophysiology and biofeedback* 40.2 (2015), pp. 107–115 (cit. on p. 38).

[98] Z. Lin et al. "Aromacue-a scent toolkit to cope with stress using the 4-7-8 breathing method". In: *Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction*. 2020, pp. 265–272 (cit. on p. 38).

[99] J. Vierra, O. Boonla, and P. Prasertsri. "Effects of sleep deprivation and 4-7-8 breathing control on heart rate variability, blood pressure, blood glucose, and endothelial function in healthy young adults". In: *Physiological Reports* 10.13 (2022), e15389 (cit. on p. 38).

[100] A. Weil. "Three breathing exercises". In: *DrWeil. com* (2017) (cit. on p. 39).

[101] K. E. Robles et al. "Aesthetics and psychological effects of fractal based design". In: *Frontiers in Psychology* (2021), p. 3413 (cit. on p. 39).

[102] R. P. Taylor. "Reduction of physiological stress using fractal art and architecture". In: *Leonardo* 39.3 (2006), pp. 245–251 (cit. on p. 39).

[103] J. Taelman et al. "Influence of mental stress on heart rate and heart rate variability". In: *4th European conference of the international federation for medical and biological engineering*. Springer. 2009, pp. 1366–1369 (cit. on pp. 46, 64).

[104] S. Boonnithi and S. Phongsuphap. "Comparison of heart rate variability measures for mental stress detection". In: *2011 Computing in Cardiology*. IEEE. 2011, pp. 85–88 (cit. on p. 46).

[105] *Inductive Respiration (RIP) Sensor Data Sheet*. PLUX– biosignalsplux. 2015 (cit. on p. 49).

[106] M. Feurer and F. Hutter. "Hyperparameter optimization". In: *Automated machine learning*. Springer, Cham, 2019, pp. 3–33 (cit. on p. 57).

[107] P. Paredes et al. "PopTherapy: Coping with stress through pop-culture". In: *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*. 2014, pp. 109–117 (cit. on p. 76).

[108] G. D. P. Regulation. "General data protection regulation (GDPR)". In: *Intersoft Consulting, Accessed in October* 24.1 (2018) (cit. on p. 77).