
UNIVERSITÀ
POLITECNICA
DELLE MARCHE

FACULTY OF ENGINEERING
MASTER'S DEGREE IN BIOMEDICAL ENGINEERING

---

# Machine Learning and Deep Learning approaches for stress detection using Empatica E4 bracelet

Supervisor
**Dr. Lorenzo Palma**

Candidate
**Ayham Altaleb**

Academic Year 2022-2023

# Acknowledgments

I would like to express my heartfelt gratitude to all those who have supported and guided me throughout my academic journey in preparing this thesis.
First and foremost, I would like to extend my deepest appreciation to my supervisor, Dr. Lorenzo Palma, for his support, invaluable guidance, and endless patience. Special thanks to Eng. Sara Campanella for her support during this journey.
I am profoundly grateful to Università Politecnica delle Marche, Italy, where I had the opportunity to obtain my master's degree. I am also thankful to Damascus University, Syria, where I received my bachelor's degree, especially to Prof. Hanan Mukhaiber, Prof. Rasha Massoud, and Prof. Hani Amasha.
My heartfelt appreciation goes out to my family for their unwavering support and encouragement: my father, Abdulrahman, my mother, Feryal, and my siblings, Bashar, Bilal, and Raghad, and my extended families, Altaleb and Bakkar; without you, I would not be here.
I would like to express my gratitude to my friends who stood by my side in Ancona, offering encouragement and understanding during the challenging times, especially to my partners during this journey, Ruba and Sameh, and to Muhammad Alkalet, Aya, Alaa, Aldreay, Omki, and Bilel.
To all those countless individuals who may not have been mentioned but have contributed in ways both big and small, I extend my sincere appreciation.
Thank you. Grazie.

*Ancona, Ottobre 2023*

Ayham Altaleb

# Abstract

In response to challenging circumstances, the human body can experience marked levels of anxiety and distress. In order to prevent stress-related complications, timely identification of stress symptoms is crucial, necessitating the need for continuous stress monitoring. Wearable devices offer a means of real-time and ongoing data collection, facilitating personalized stress monitoring. This study aimed to detect stress by analyzing physiological signals collected through the Empatica E4 bracelet. Machine Learning algorithms (Random Forest, SVM, Logistic Regression) and Deep Learning pre-trained CNNs (GoogLeNet, SqueezNet) were employed to differentiate between stressful and non-stressful situations. Data from 29 subjects, including photoplethysmographic (PPG) and electrodermal activity signals (EDA), were used to extract 27 features with and without overlapping. These features were then utilized in three Machine Learning algorithms for binary classification using Python, after applying the Chi-square test and Pearson's correlation coefficient via WEKA for feature importance ranking. Additionally, SHapley eXplainable AI was applied to the top-performing model, Random Forest, in the overlapping case, shedding light on the most impactful features and comparing them with feature selection methods. Notably, HRV (Heart Rate Variability) features emerged as significant in stress detection. Furthermore, in the non-overlapping case, continuous wavelet transform was applied to PPG signals to generate scalograms, which were subsequently fed into two different pre-trained CNNs. The study's results showcased the overlapping had a positive impact on all models. Moreover, the Random Forest model is the highest-performing, achieving an accuracy of 76.4% without overlapping and an impressive 99.5% with overlapping segments. Additionally, Deep Learning models exhibited potential in stress classification, particularly when considering the use of PPG signals only.

# Contents

# List of Figures

# List of Tables

# Introduction

Stress is a pervasive aspect of modern life, and its impact on both physical and mental well-being cannot be understated. In response to challenging circumstances, the human body can experience marked levels of anxiety and distress, which, if left unmanaged, can lead to a range of stress-related complications. Timely identification of stress symptoms is crucial for effective stress management and prevention, necessitating the need for continuous stress monitoring. Wearable devices have emerged as a promising solution for real-time and ongoing data collection, enabling personalized stress monitoring.
This thesis aimed at stress detection and monitoring through the analysis of physiological signals, focusing on data collected using the Empatica E4 bracelet. The objective of this study is to utilize both traditional Machine Learning algorithms and cutting-edge Deep Learning techniques to differentiate between stressful and non-stressful situations based on these physiological signals.
In this introduction, we provide a roadmap of what to expect within the pages of this thesis:

**Chapter 1- Anatomy and Physiology of the Human Heart and Skin :** To understand the cardiovascular system and the skin.

**Chapter 2- Stress, Bio-Signals, and Wearable Sensors:** This chapter explores core concepts of stress and the various methods employed for stress assessment. Moreover, introduces the fundamental role of wearable sensors in real-time data collection, which is crucial for personalized stress monitoring. Key bio-signals, including the electrocardiogram (ECG), photoplethysmographic (PPG) signals, and electrodermal activity (EDA), are explored in the context of stress assessment.

**Chapter 3- Artificial Intelligence:** Machine Learning and Deep Learning are at the core of this thesis. We explore the fundamentals of these techniques, feature selection, data splitting methods, and a range of classifiers, including Support Vector Machine, Logistic Regression, Decision Trees, Random Forest, and Artificial Neural Networks. Additionally, Deep Learning and the significance of hyperparameter tuning are introduced, alongside the concept of Explainable AI.

**Chapter 4- Literature Review:** This chapter provides an overview of related research in the field of stress detection and monitoring. We discuss the methodologies and results of various studies to provide context for our own investigation.

**Chapter 5- Materials and Methods:** Here, we detail the materials and data acquisition protocol, including the use of the Empatica E4 bracelet. We also present the machine learning and deep learning approaches employed in this study, outlining the methodologies for stress classification.

**Chapter 6- Results:** This chapter presents the findings of our study, focusing on feature selection, the performance of machine learning approaches, and the efficacy of deep learning models in stress classification.

**Chapter 7- Discussion:** The discussion chapter provides a critical analysis of the results, exploring the significance of feature selection, the performance of machine learning and deep learning models, and model explainability.

**Chapter 8- Conclusions:** The thesis culminates in a concise summary of the study's key findings, offering insights into the potential of physiological signals and wearable technology in stress detection and monitoring.

Through a comprehensive examination of these chapters, this thesis aims to contribute to the growing body of knowledge in the field of stress assessment, with a particular focus on the application of machine learning and deep learning techniques for accurate and timely stress detection.

# Chapter 1

# Anatomy and physiology of the human heart and skin

## 1.1 Introduction

Stress is a common experience that can have negative impacts on mental and physical health. Accurately detecting and monitoring stress is crucial for preventing these negative outcomes. Bio-signals, such as electrocardiogram (ECG), photoplethysmography (PPG), and electrodermal activity (EDA), have been shown to be reliable indicators of stress. Wearable sensors have made it easier to collect these bio-signals in real time, which has opened up new possibilities for stress detection and management. However, understanding the anatomy and physiology underlying these bio-signals is essential for accurately interpreting the data collected from wearable sensors. This thesis aims to explore the relationship between bio-signals and stress using machine learning, with a particular focus on the role of anatomy and physiology in stress detection. By gaining a deeper understanding of the physiological processes involved in generating bio-signals, we can improve the accuracy of stress detection and develop more effective interventions for stress management.

## 1.2 Cardiovascular system

The cardiovascular system consists of the heart and blood vessels, while the lymphatic system collects excess fluid from the tissue interstitium and returns it to the venous circulation. The heart can be viewed as two pumps with the pulmonary and systemic circulations in between. Pulmonary circulation involves the exchange of gases between the blood and alveoli in the lungs, while systemic circulation comprises all blood vessels within and outside of organs except for the lungs. The right side of the heart receives venous blood from the systemic circulation and pumps it into the pulmonary circulation for gas exchange, while the left side receives oxygenated blood from the lungs and ejects it into the aorta for distribution to all organs via the arterial system. The capillaries within the organs are the primary site of exchange. Blood flow from the capillaries enters veins, which return blood flow to

the right atrium via large systemic veins.
The cardiovascular system is arranged in series, with the right and left sides of the heart separated by the pulmonary and systemic circulations, and in parallel shown in Figure 1.1, with most of the major organ systems receiving their blood from the aorta and returning it to the heart via the venous system[1][2][3].

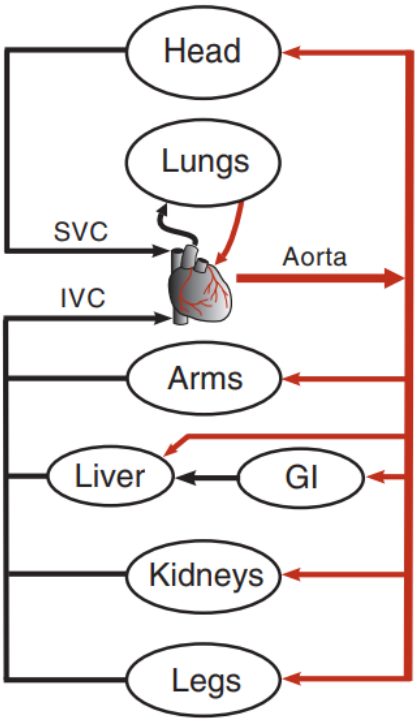



Figure 1.1: Parallel arrangement of organs within the body. GI, gastrointestinal circulation. SVC, superior vena cava; IVC, inferior vena cava [2].

### 1.2.1 The heart

The heart is the key organ of the cardiovascular system and is responsible for maintaining continuous blood flow. The wall of the heart is composed of three layers: the epicardium (the outer layer), the myocardium (the middle layer), and the endocardium (the inner layer). The myocardium is the muscle of the heart, accountable for its pumping action. The heart cavity is made up of two portions, left and right, separated by an inner wall called a septum. There are four valves in the human heart, two are between heart chambers called "atrioventricular valves", and two between chambers and vessels called "semilunar valves". On the right side, the atrium and ventricle are divided by the tricuspid valve, and on the left side are divided by the mitral valve. Furthermore, the pulmonary valve is located between the right ventricle and the pulmonary artery, and the aortic valve is located between the left ventricle and the aorta. Figure 1.2 shows the simplified structure of the heart [4][2][1].

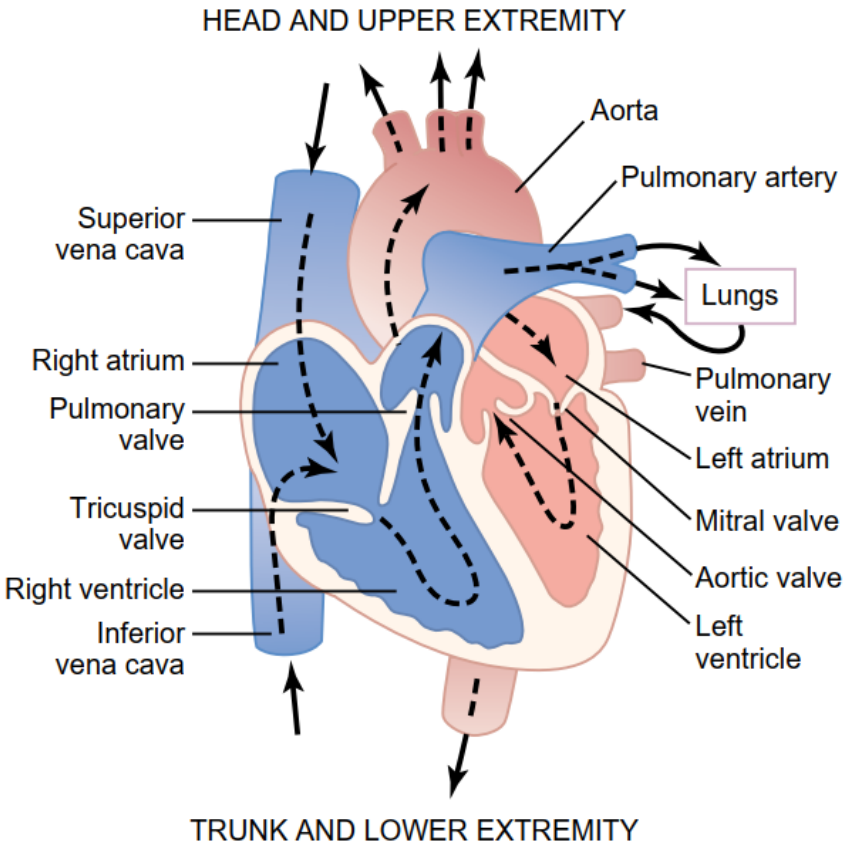



Figure 1.2: Structure of the heart [1].

### 1.2.2 Conduction system

The heart possesses the ability to generate self-generated electrical impulses and regulate the pathway of these impulses through a distinctive conduction system. This system is composed of five components, including:

- Sinoatrial node (SA).
- Atrioventricular node (AV).
- Bundle of His.
- The left and right bundle branches.
- The Purkinje fibers.

The sinoatrial node is a specialized cardiac muscle that is small, flattened, and ellipsoid in shape, located in the superior posterolateral wall of the right atrium. The atrioventricular node is found in the posterior wall of the right atrium, subdivided into the lower nodal bundle and compact node. The bundle of His consists of specialized muscular tissue responsible for electrical conduction, which further divides into two branches to transmit impulses to the left and right ventricles. Lastly, the Purkinje fibers are located beneath the endocardium in the inner ventricular

walls of the heart. The positions of these structures within the heart are illustrated in Figure 1.3. Conduction velocities of different regions are noted in parentheses [1].

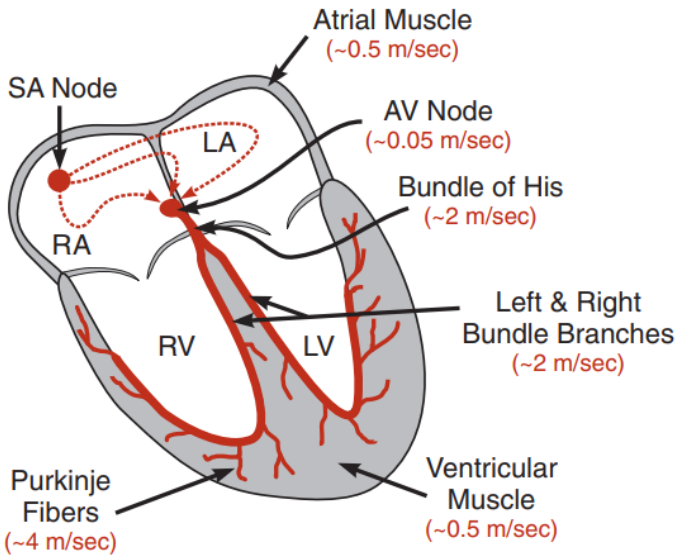



Figure 1.3: Conduction system within the heart [2].

### 1.2.3 Cardiac cycle

The cardiac cycle encompasses the events that occur from the beginning of one heartbeat to the start of the next. The cycle begins with the spontaneous generation of an action potential in the sinus node. This impulse travels through both atria and into the ventricles via the A-V bundle, with a delay of more than 0.1 seconds to allow for atrial contraction before a ventricular contraction begins. The cardiac cycle comprises diastole, a relaxation period during which the heart fills with blood, followed by systole, a contraction period. The mechanical events during the cardiac cycle can be summarized as follows:

- The cycle begins with an almost synchronous contraction of the two ventricles, resulting in a rapid increase in blood pressure in the ventricles.
- As the pressure in the ventricles exceeds that in the atria, the mitral valve closes, producing the first heart sound and marking the start of systole.
- During the isovolumetric phase, there is no change in ventricular volume as the pressure in the ventricles continues to rise until it exceeds that in the aorta, causing the aortic valve to open and blood ejection into the systemic circulation to begin.
- As the ventricular wall tension decreases, the pressure gradient between the ventricles and the aorta reverses, and flow decelerates until the aortic valve closes, generating the second heart sound and marking the onset of diastole.

- During the second isovolumetric period, the ventricular muscle relaxes, and the pressure in the ventricles decreases, while the pressure in the atria rises as the left atrium is filled by the pulmonary venous system.
- As the pressure in the atria exceeds that in the ventricles, the mitral valve reopens, allowing the ventricles to refill with blood. This process initially occurs passively, driven by a pressure difference between the atria and ventricles, and then becomes active as the atria contract during atrial systole, pushing the remaining of blood volume.
- Shortly after, the ventricles contract again, restarting the cycle.

Figure 1.4 illustrates the different events during the cardiac cycle on the left side of the heart, including pressure changes in the aorta, left ventricle, and left atrium, as well as changes in left ventricular volume, the electrocardiogram, and a phonocardiogram. The electrocardiogram displays P, Q, R, S, and T waves, representing electrical voltages generated by the heart and recorded by the electrocardiograph. The P wave is caused by atrial depolarization, followed by atrial contraction, while the QRS waves indicate ventricular depolarization and contraction. The T wave represents ventricular repolarization and occurs just before the end of ventricular contraction [1][2].

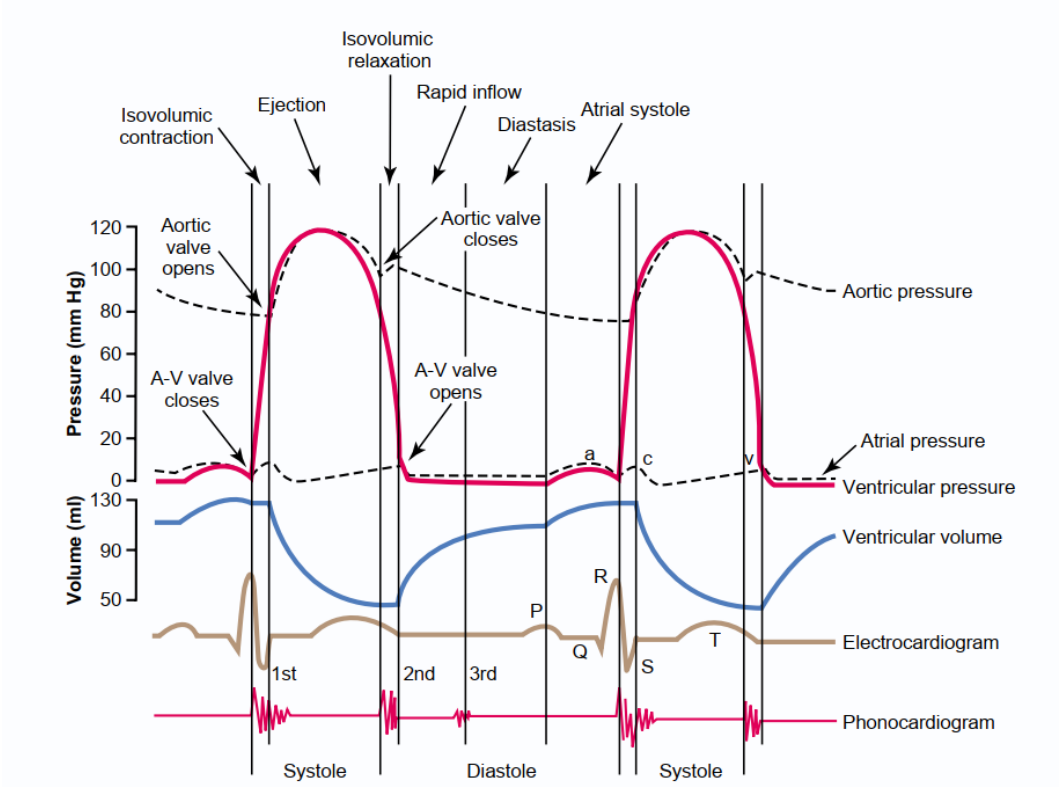



Figure 1.4: Events of the cardiac cycle for left ventricular function [1].

### 1.2.4 Circulatory system

The circulation system functions to transport nutrients, waste products, and hormones, and maintain an optimal environment for the body's tissues. Arteries transport blood under high pressure, arterioles control blood flow into capillaries, capillaries exchange substances between the blood and interstitial fluid, venules collect blood from capillaries, and veins transport blood back to the heart while also serving as a reservoir for extra blood shown in Figure 1.5. Venous walls are thin but muscular enough to contract or expand depending on the needs of the circulation. When

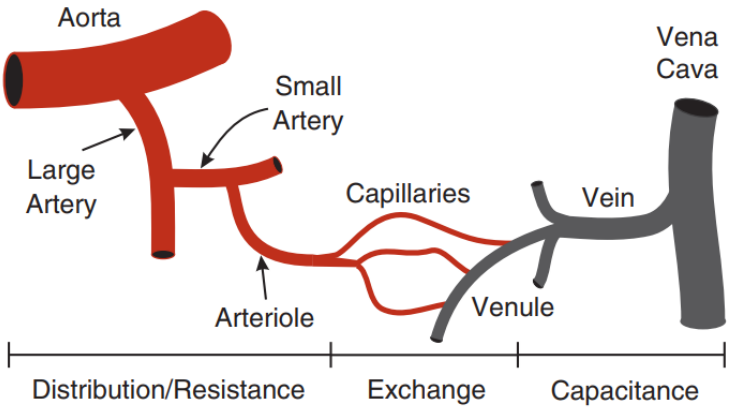



Figure 1.5: Major types of blood vessels found within the circulation [2].

the left ventricle ejects blood into the aorta, a characteristic aortic pressure pulse is produced, consisting of a peak systolic pressure, a notch (dicrotic notch or incisura), a small increase in pressure (dicrotic wave), and a diastolic pressure. The difference between systolic and diastolic pressures is the aortic pulse pressure. Mean arterial pressure, which is the average pressure over time, is the primary pressure that drives blood flow in organs, and it needs to be determined when assessing vascular function. Figure 1.6shows pressure pulse within the aorta.[1][2].

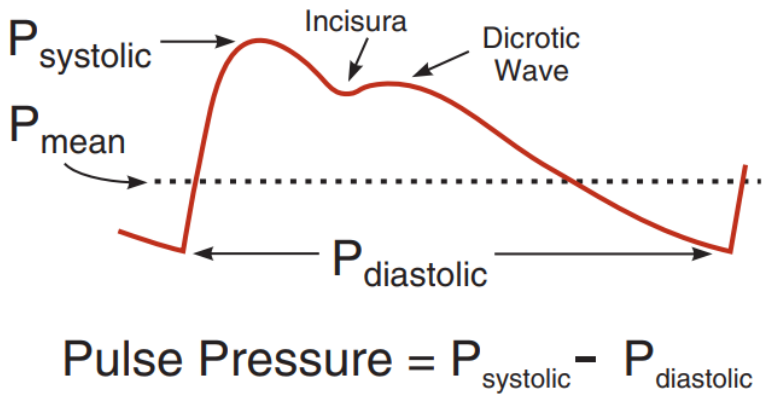



Figure 1.6: Pressure pulse within the aorta [2].

### 1.2.5 Neurohumoral control of the heart and circulation

The parasympathetic and sympathetic nervous systems are two branches of the autonomic nervous system that have opposing effects on the heart and circulatory system. The parasympathetic nervous system acts through the release of acetylcholine, which binds to muscarinic receptors in the heart and causes a decrease in heart rate. In contrast, the sympathetic nervous system acts through the release of norepinephrine, which binds to beta-adrenergic receptors in the heart and causes an increase in heart rate and contractility. The interaction between the two systems helps to maintain a balance in heart rate and blood pressure. In a healthy individual, the parasympathetic and sympathetic nervous systems work together to maintain a stable heart rate and blood pressure. However, in certain situations such as exercise or stress, the sympathetic nervous system becomes dominant, leading to an increase in heart rate and blood pressure. The sympathetic nervous system also has effects on the circulatory system, including vasoconstriction and vasodilation. When the sympathetic nervous system is activated, it causes vasoconstriction of blood vessels in non-essential organs such as the digestive system, while simultaneously causing vasodilation of blood vessels in essential organs such as the heart and brain. This redistribution of blood flow helps to ensure that these vital organs receive the necessary nutrients and oxygen. Furthermore, the sympathetic nervous system also stimulates the release of epinephrine and norepinephrine from the adrenal medulla, which have effects on the heart and circulation. Epinephrine and norepinephrine increase heart rate, contractility, and vasoconstriction, leading to an increase in blood pressure and cardiac output. Table 1.1 shows the effects of sympathetic and parasympathetic stimulation on cardiac a vascular function and Figure 1.7 shows the organization of sympathetic and vagal innervation of the heart and circulation [2].

Table 1.1: Effects of sympathetic and parasympathetic stimulation on cardiac a vascular function [2].

| | SYMPATHETIC | PARASYMPATHETIC |
|---|---|---|
| **Heart** | | |
| Chronotropy (rate) | + + + | - - - |
| Inotropy (contractility) | + + + | - |
| Dromotropy (conduction velocity) | + + | - - - |
| **Vessels (Vasoconstriction)** | | |
| Resistance (arteries, arterioles) | + + + | - |
| Capacitance (veins, venules) | + + + | 0 |

Relative magnitude of responses (+, increase; -, decrease; 0, no response) indicated by number of + or - signs.

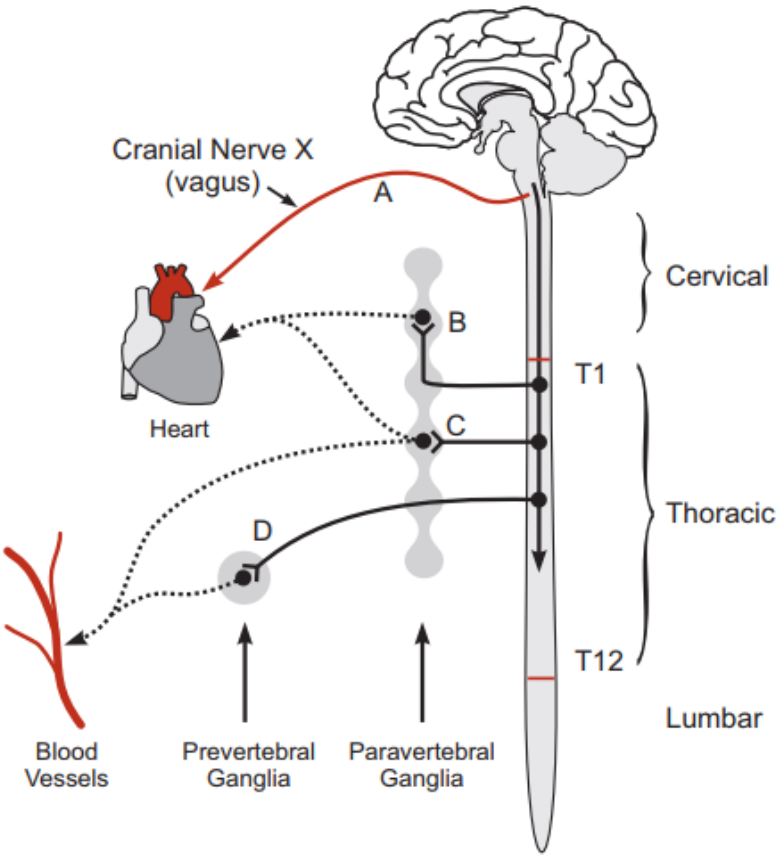



Figure 1.7: Organization of sympathetic and vagal innervation of the heart and circulation [2].

## 1.3 Skin

To interpret skin conductance and potential, it's important to understand the structure of tissues both on and beneath the skin surface. Figure 1.8 illustrates the key features of the skin, including the epidermis, which is the most superficial layer consisting of the stratum corneum, stratum lucidum, granular layer, prickle cell layer, and basal/germinating layer. The corneum layer is made up of dead cells at the skin's surface, with healthy living cells found at its base, and transitional cells in between. This layer is also referred to as the horny layer. The dermis contains blood vessels, while the eccrine sweat gland secretory cells are found at the border of the dermis and panniculus adiposus or superficial fascia. The eccrine sweat gland excretory duct is a simple tube made of epithelial cells, which ascends and opens on the skin's surface. Cholinergic stimulation via fibers from the sympathetic nervous system is the primary influence on the production of sweat by these glands.
The epidermis normally has a high electrical resistance due to the thick layer of dead cells with thickened keratin membranes, which is expected as the skin functions as a barrier against external factors like abrasion and mechanical assaults. Experiments show that the entire epidermis, except for desquamating cells, is a permeability barrier to flow and behaves as a passive membrane. However, sweat ducts penetrate

the corneum layer from underlying cells, resulting in a relatively good conductor as sweat is a weak electrolyte with many low-resistance parallel pathways[5].

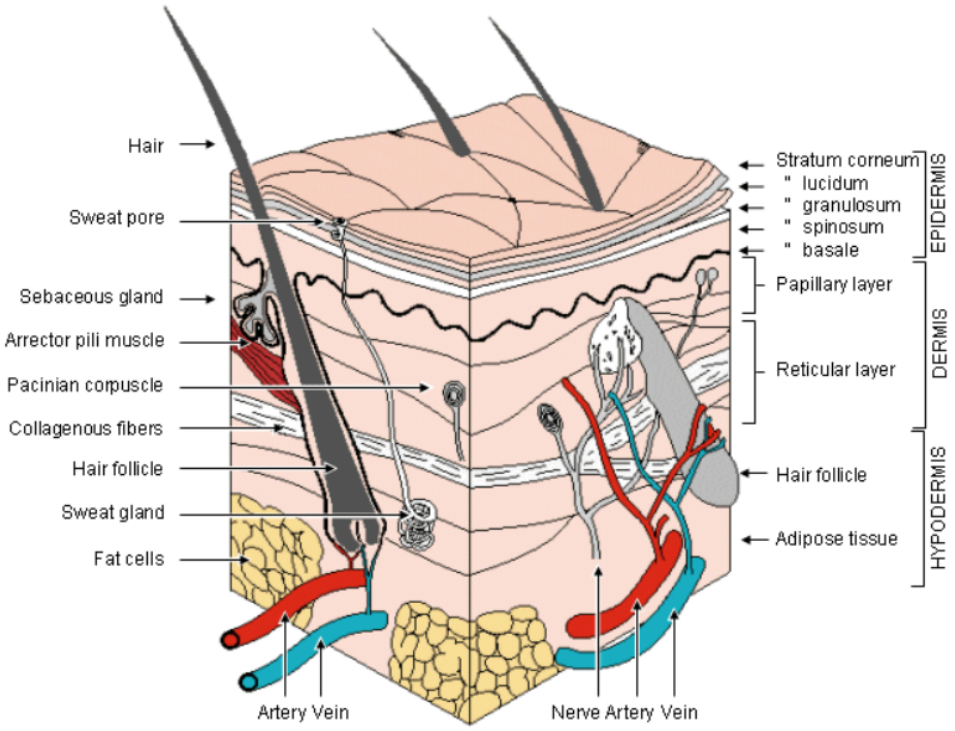



Figure 1.8: Section of smooth skin taken from the sole of the foot [5].

# Chapter 2

# Stress, Bio-Signals, and Wearable Sensors

## 2.1 Introduction

One of the main factors contributing to both physical and mental illnesses in people is stress [6]. An organism's natural reaction to an intrinsic or extrinsic situation, whether it be favourable or unfavourable, physical or mental, is known as stress [7]. It is the body's method of coping with an oppressive or negative situation and constantly works to restore the body to its normal balance [8]. Stress-related pathologies or disorders are thought to be the second most common cause of disease in both Europe and the United States, accounting for three out of every four doctor visits. [9]. The first stage of stress is the disruption of an organism by a stimulus or event known as stressors [8]. Although stressors can take on many different forms, they can be broadly divided into two categories: psychological and physiological. Psychological stressors include things like debt, the death of a loved one, losing a job, studying for an exam, and other similar items. Physiological stressors include things like infections, high temperatures, and a lack of relaxation. When the body perceives a situation as stressful, it can trigger short-term or long-term reactions. The hypothalamus in the brain plays a crucial role in this process by activating and sending signals to the pituitary gland, which then stimulates the adrenal gland to produce cortisol. This hormone helps to stabilize the blood sugar supply and restore the body to normal function. In addition, the adrenal medulla, which is part of the autonomic nervous system, is stimulated by the hypothalamus to produce short-term stress responses. This results in the release of adrenaline, which causes the fight-or-flight response and activates the sympathetic nervous system. Once the stressor is removed and the parasympathetic nervous system takes over, the body returns to its normal state [10].

Based on the time-lapse, stress can be divided into three categories and each of them has a unique set of symptoms, traits, duration, and treatment options. It is distinguished into acute stress, the most common, characterized by short duration and associated with negative thoughts, episodic stress, which happens when intense

stress is sustained over a long period before it becomes a habit, and chronic stress, which might be the result of early childhood experiences and traumatic experiences from the past that have shaped one's life [11].

## 2.2 Stress

Stress is a condition where the mind and body react to external and internal factors, leading to anxiety, depression, and tension. Work-related stress is consistently associated with negative effects on cognitive and mental health, workplace performance, and an increased risk of disease. These consequences are often linked to changes in brain functions, particularly in areas like the hippocampus, prefrontal cortex, and amygdala. Quality of life (QoL) is affected by stress, and it has negative associations with health, including the immune, digestive, nervous, and cardiovascular systems. Chronic stress can lead to brain changes, memory problems, and impaired learning. It also weakens the immune system and can contribute to heart problems. In summary, stress has significant effects on the human biological system. [12] [13] Figure 2.1 shows the Impact of stress on main brain areas, cognitive functions, and affective domains. And Figure2.2 shows Stress impacts on human immune system, digestive

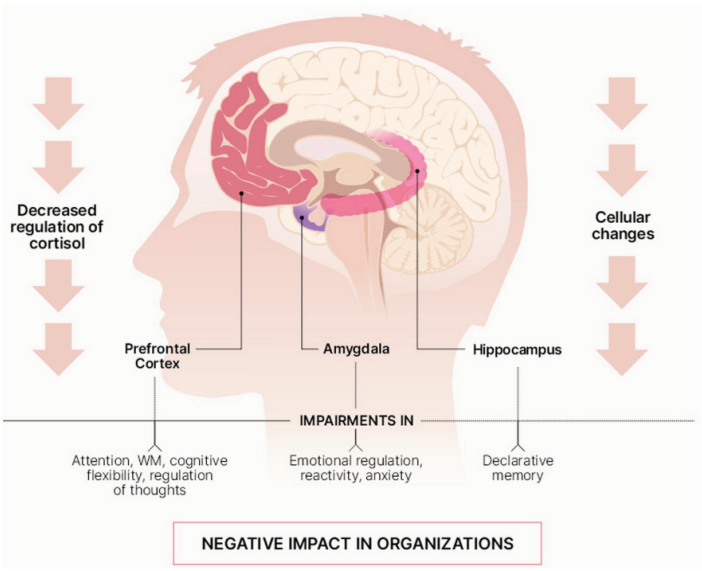



Figure 2.1: Impact of stress on main brain areas, cognitive functions, and affective domains [14].

system, central nervous system, and cardiovascular system, which are summarized as follows:

• **Brain:** Stress affects neural pathways and cognitive processes, particularly memory and cognition. It triggers changes in brain chemistry, involving dopamine, noradrenaline, and glucocorticoids, which impact memory consolidation and retrieval. Stress also influences cognition, with effects dependent on factors like duration and intensity, potentially leading to cognitive disorders and alterations in brain function.

• **Cardiovascular system:** Psychological stress is a recognized risk factor for cardiovascular diseases, affecting heart rate and blood pressure. Stress activates the sympathetic nervous system, leading to vasoconstriction, increased blood pressure, blood clotting disorders, and vascular changes, all contributing to cardiac arrhythmias and heart attacks. Chronic stress in personal life is associated with a significant increase in coronary heart disease development. High cortisol levels from long-term stress can raise cholesterol and triglycerides, promoting heart disease. Stress also disrupts blood pressure, potentially leading to arterial plaque buildup.
• **Digestive system:** The intestinal nervous system, with around 100 million nerve cells, operates in the gastrointestinal area and connects bidirectionally with the central nervous system through the sympathetic and parasympathetic nervous systems. Stress worsens symptoms of gastroesophageal reflux disorder (GERD), particularly in individuals with high gastrointestinal susceptibility. Stress is a risk factor for upper gastrointestinal diseases, such as peptic ulcers and inflammatory bowel disease, and major stressors can influence disease activity. In summary, stress significantly affects various gastrointestinal disorders and their symptoms.[12]

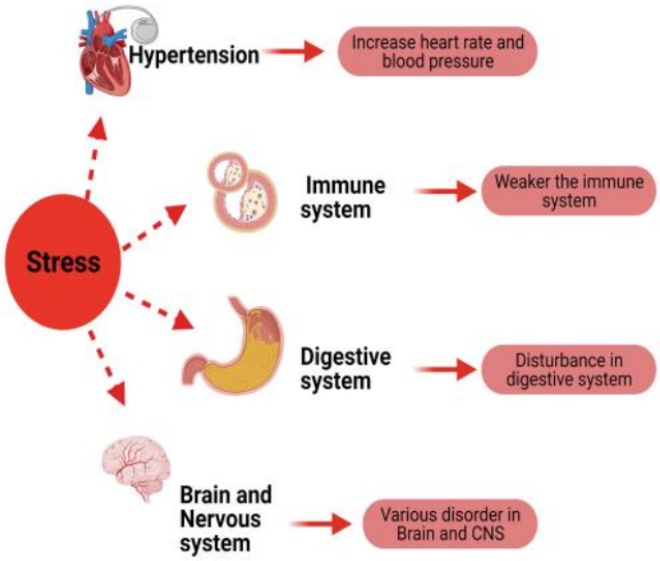



Figure 2.2: Stress impacts on human immune system, digestive system, central nervous system, and cardiovascular system [12].

Stress is a heterogeneous disease that affects adults and young people. Due to the demanding physical and mental efforts required of employees, the workplace has become a major source of stress in the latest days [15]. It could also be a result of staff not having the resources they require to do their jobs well or of staff not having their needs met. Stress at work has been linked to frequent absences, mistakes, and lower productivity [16]. According to evidence, the EU spends about €617 billion a year on social welfare, health care, and programs to help people who are stressed out or depressed at work [17]. This demonstrates how stress at work not only affects the productivity of individuals but also the entire state. Teenagers frequently experience

academic stress, a type of mental distress brought on by the many expectations that are placed on them. It can be difficult to avoid stress as a factor. Students experience stress due to a variety of demands, including homework, exams, classes, projects, friends, and family. Their academic success is directly correlated with these demands. Students under high stress often experience depression and anxiety [18].

### 2.2.1 Stress Assessment tests

Stress can be evaluated either **1)** subjectively through structured scales, questionnaires, or surveys. Although they are inexpensive and simple to use, questionnaires have some drawbacks that make them less useful since they are based on individual perceptions [19] . Or **2)** objectively by measuring physiological responses. Common clinical stress assessment tools involve self-reported questionnaires like Cohen's Perceived Stress Scale (PSS) and visual scales such as the Visual Analogue Scale for Stress (VASS). Biomedical researchers often prefer biochemical markers like cortisol and alpha-amylase and induce stress using tests like the Trier Social Stress Test (TSST). Alternatively, some studies assess stress by monitoring the body's physiological signals[20]. Studies have revealed that, in addition to the conventional methods of detecting stress through questionnaires and behavioural observations, it can also be determined and measured from physiological, psychological, and neurological responses [21]. Below is a summary of commonly used stress assessment methods, outlined in Table 2.1

Table 2.1: Stress assessment tests and brief detail [20].

| **Test Name** | **Stress Assessment Method** |
|---|---|
| Mental Arithmetic Test | To create stress, participants are given a time limit and asked to complete mathematical problems (subtraction, multiplication). |
| Trier Social Stress Test (TSST) | Requires participants to deliver speeches on predetermined topics with little advance notice. The participants are also required to carry out some verbal calculations after the speech. Both tasks are carried out in front of an audience that will be giving feedback. |
| Stroop Test | Instead of reading the words, participants are presented the names of several colors printed in different font colors and asked to identify the font color. |

*Continued on next page*

Table 2.1 – *Continued from previous page. Stress assessment tests and brief detail.*

| **Test Name** | **Stress Assessment Method** |
|---|---|
| Perceived Stress Scale (PSS) | Participants fill out the questionnaire by rating the questions about their feelings and thoughts. The total score varies from 0 (no stress) to 40 (highest stress). |
| Visual Analogue Scale for Stress (VASS) | In this test, rather than providing a numerical response for each question, participants are asked to rate their level of stress on a scale as no stress, moderate stress, or high stress. Most of the time, a 5-point (smiley) scale is used for stress assessment. |
| Stress Response Inventory (SRI) | The Stress Response Inventory consists of 39 questions scored in the range of 0 to 156. Tension, weariness, despair, aggression, anger, somatization, and frustration are the 7 components that make up these questions. High perceived stress corresponds to a high score. |
| COPE Inventory | The purpose of the 28 self-reporting questions is to gauge how effectively individuals deal with stressful situations. On a scale from 1 (low stress) to 4 (high stress), a score is assigned to each question. The final score identifies the individuals' approach or avoidant stress coping strategy. |
| Holmes and Rahe Stress Inventory | Calculates the level of stress experienced over the last year. From a list of 43 life events related to stress, participants choose those that took place in their lives. Scores for each event vary. Participants who reach a score of more than 300 have a larger risk of being unwell, whereas a score of less than 150 indicates a modest risk. |
| State-Trait Anxiety Inventory (STAI) | Twenty items that assess trait and actual anxiety are validated by the participants. Participants answer the questions on a scale from 1 to 4, with 1 signifying the least stress and 4 signifying a high level of stress. |

*Continued on next page*

Table 2.1 – *Continued from previous page. Stress assessment tests and brief detail.*

| Test Name | Stress Assessment Method |
|---|---|
| Montreal Imaging Stress Task (MIST) | The three phases of MIST are rest, control, and experiment. The participant stares at the computer's static screen while they are resting. The subject is given a series of mathematical problems to solve in the control stage, while in the experiment stage, challenging and time-limited arithmetic assignments are presented to induce high stress. |
| Perceived Stress Questionnaire (PSQ) | Participants fill out two types of questionnaires consisting of 30 questions; the first questionnaire has questions about stressful events and emotions in the previous 2 years while the second one has questions about stress during the previous month. Each question must receive a score from 1 (no stress) to 4 (stress). |

## 2.3 Wearable sensors

In today's digital era, the term "wearable" has taken on a new meaning. It no longer refers to traditional clothing but rather to accessories with functionality and mobile information processing capabilities. These wearables include smartwatches, head-mounted displays, sensors, and smart garments. They have expanded beyond fashion and protection to provide personalized and configurable mobile information processing for various applications such as gaming, fitness, healthcare, and entertainment. The use of wearables has transformed various aspects of our lives. Wearables are also employed in critical areas such as public safety, where they help monitor the physical condition of first responders and detect hazardous materials. Furthermore, wearables are utilized in monitoring racecar drivers' health and enhancing the viewing experience for fans. The value of wearables extends to diverse user groups. For instance, they assist the "sandwich generation" in caring for elderly parents by monitoring their health and promoting independence. Wearables have also been employed in parenting, allowing parents to monitor the well-being of young children. Fundamentally, wearables perform basic functions such as:

- Sensing.
- Processing.
- Storing.
- Transmitting.

- Applying data.

Figure 2.3 shows a schematic representation of the unit operations associated with obtaining and processing situational data using wearables. For example, if dangerous gases are detected by a wearable on a first responder, the data can be processed in the wearable and an alert issued. The specific operations depend on the application domain and the wearer. Processing may occur either on the individual or at a remote location. Wearables play a crucial role in obtaining and processing situational data, enabling real-time alerts, confirmatory testing, and personalized responses.
While wearables find applications in various fields, this chapter focuses primarily on their role in healthcare. Wearables offer a non-intrusive means of longitudinally monitoring individuals, aiding in the early detection of problems and diseases for preemptive care and improved quality of life. The principles and concepts discussed in the healthcare domain can be applied to other application areas with ease. [22]

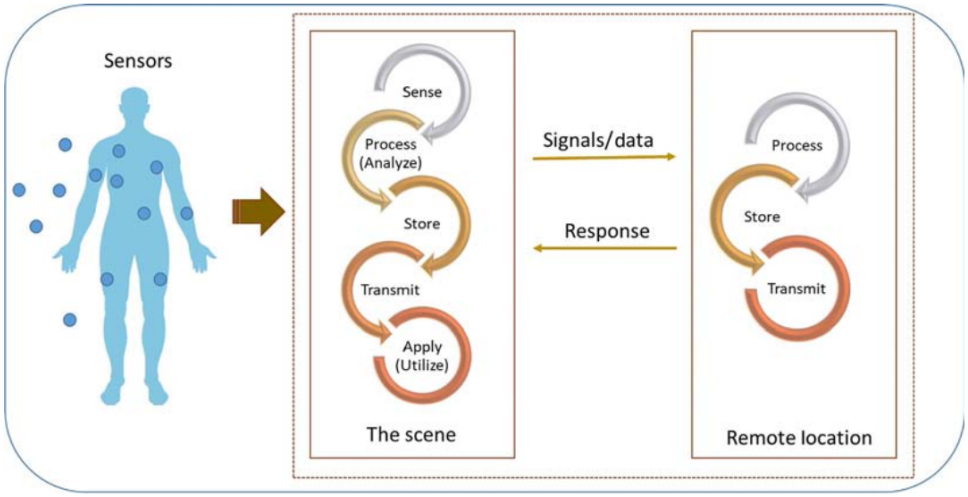



Figure 2.3: Unit operations in obtaining situational awareness: the role of wearables [22].

Smart wearable devices that can measure signals even in natural settings for assessing cognitive and sensory states have been made possible by recent advancements in embedded systems and sensors. Nowadays, vital signals are collected by means of several variegate wearable devices - smart watches, chest belts, smart t-shirts, and head-mounted devices [23]- allowing ongoing mental health monitoring to be easier compared to the past. The widespread market adoption of smart wearables has given people the ability to track, store, and transfer personal information about their surroundings, physical activity, and health [24].

## 2.4 Bio-signals and mental stress correlation

Research has consistently shown an association between higher heart rate and stress. This alteration in heart rate influences blood flow within the body, which can be

monitored using an electrocardiograph (ECG) signal. Blood flow changes can be measured through blood volume pulse (BVP) derived from a photoplethysmography (PPG) signal. Sweat release during stress affects skin conductance, measured by an electrodermal activity (EDA) measurement device, which is also known as galvanic skin response (GSR). Muscle tension, linked to stress, is monitored using electromyography (EMG). Brain signals (EEG) are also indicative parameters because they are connected to the autonomic nervous system. Additionally, skin temperature (ST) and accelerometer (ACC) sensors can aid in stress detection, as chronic stress may also lead to mild fever, anxiety, and restlessness[20] [25] [26]. Figure 2.4 shows common places of wearable sensors on the human body.

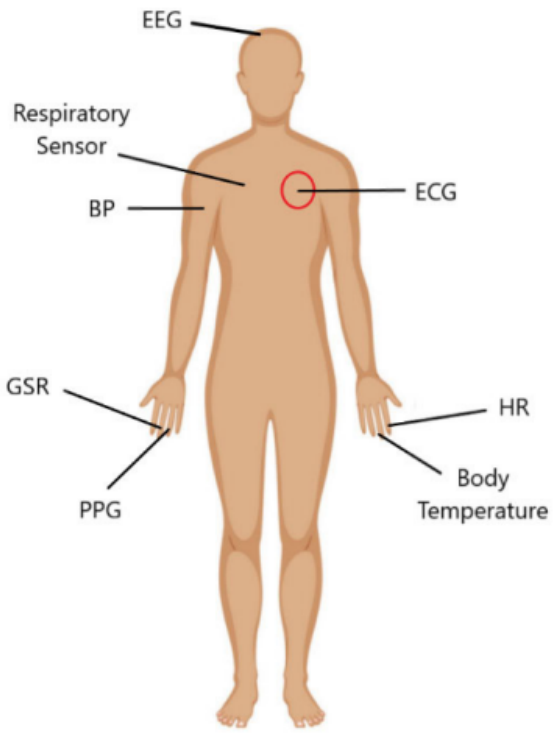



Figure 2.4: Schematic diagram showing common places of wearable sensors on human body [25].

### 2.4.1 Electrocardiogram (ECG)

Cardiovascular parameters are extremely useful to investigate the human condition. The ECG is one of the most common heart tests used in the assessment and diagnosis of Cardiovascular diseases. As the cells in the heart undergo depolarization and repolarization, electrical currents are generated and conducted through the surrounding tissues, spreading throughout the body. These electrical currents can be measured using a set of electrodes placed on specific locations on the body surface, and the resulting recording is called an electrocardiogram (ECG). The ECG consists of repeating

waves that represent the sequence of depolarization and repolarization in the atria and ventricles. The ECG does not measure absolute voltages but instead captures changes in voltage relative to a baseline level. Typically, ECGs are recorded on paper at a speed of 25 mm/s, with a vertical calibration of 1 mV/cm. The ECG intervals can provide information about the rate, the rhythm, and the electrical activity of the heart. Figure 2.5 represents an ECG typical waveform, recorded by placing electrodes on the surface of the human body and characterized by:

- The P-wave: it represents the depolarisation of the atria.
- The QRS complex: it represents the depolarisation of the ventricles (the most prominent wave in ECG).
- The T-wave: it represents the repolarization of the ventricles.
- The PR interval: it represents the conduction of the impulse from the atrium to the ventricles.
- The ST segment: it represents the beginning of the ventricular repolarization. [2]

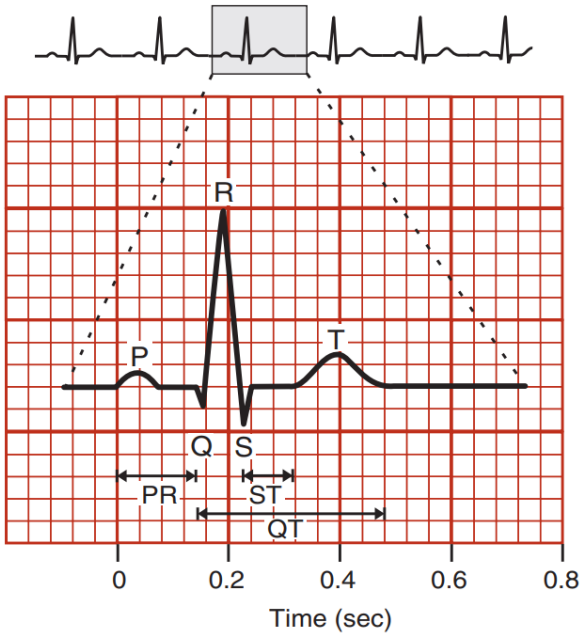



Figure 2.5: Components of the ECG trace [2].

### 2.4.2 PPG signal

The PPG waveform is shown in Figure 2.6, obtained by measuring the amount of light absorption after transmission through or reflection from human tissue, consisting of a pulsatile component and a non-pulsatile component. The pulsatile component,

known as the AC component, is synchronized with the cardiac cycle and reflects changes in blood volume in the artery. It is influenced by factors such as vasodilation, vasomotor activity, and vascular tones. The non-pulsatile component, called the DC component, includes all other components of the PPG waveform and is affected by biological characteristics, external factors, and physiological activities.
The amplitude of the PPG waveform varies among individuals due to factors like physical characteristics, tissue composition, and blood vessel distribution. The measurement of the PPG waveform is also influenced by environmental factors such as ambient light. The PPG waveform changes not only with cardiac activity but also due to respiration, autonomic nervous system activity, arterial activity, and venous activity Figure 2.7 shows the main three families of factors that influence PPG signal. The PPG waveform exhibits a rising curve during systole (cardiac contraction) and a descending curve during diastole (cardiac dilation). Feature points in the waveform include pulse onset, systolic peak, dicrotic notch, and diastolic peak. The absolute value of PPG amplitude cannot be directly compared across individuals due to variations in body tissues and individual characteristics. PPG baseline is affected by various factors such as respiration, vascular compliance, vascular tone, pain, and drug use. The amplitude of the systolic peak, a representative characteristic of the PPG waveform, is correlated with microvascular expansion and cardiac output. Changes in vascular tone and compliance affect the occurrence of the dicrotic notch. Aging influences the time difference between the diastolic peak and the systolic peak. [27]

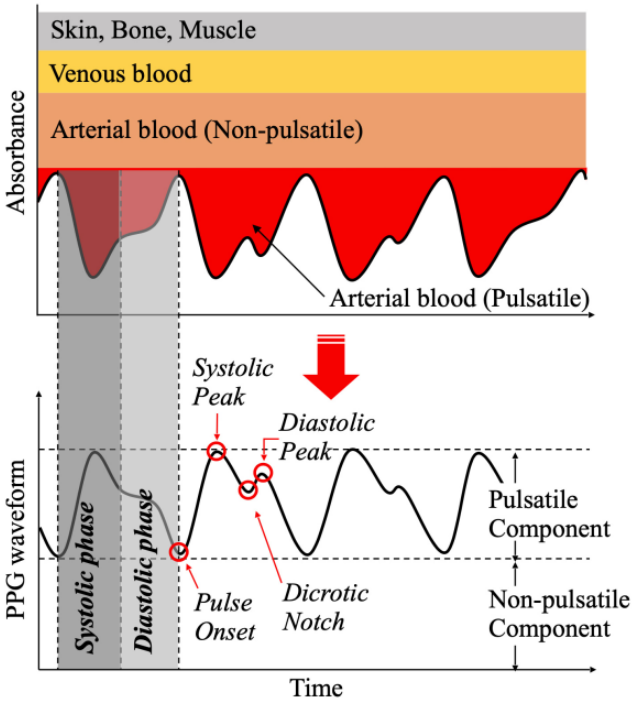



Figure 2.6: Principle of phototoplethysmogram generation and waveform features [27].

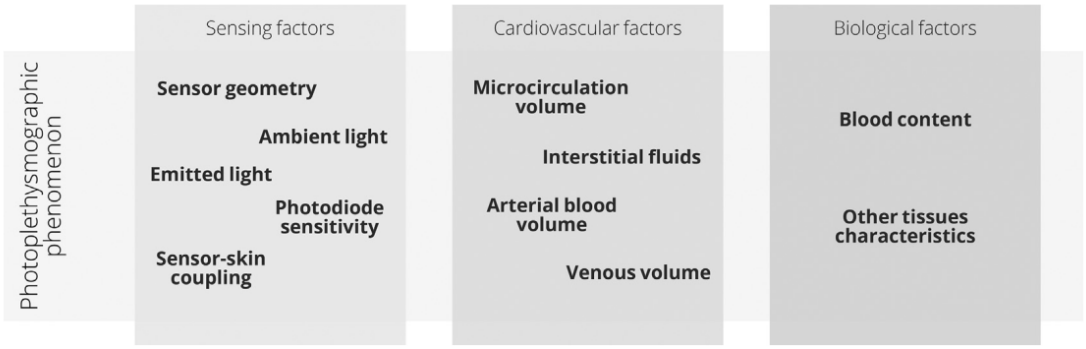



Figure 2.7: An overview model of the PPG phenomena and its three families of factors that influence PPG signal [22].

### 2.4.3 Parameters derived from ECG and PPG signals

Heart rate variability (HRV) is a valuable measure of the parasympathetic nervous system's influence on heart activity, commonly used in psychophysiology and cardiovascular research. It offers a non-invasive and cost-effective way to assess cardiac function. Traditional HRV assessment relies on electrocardiogram (ECG) recordings, which provide high accuracy but require electrodes and may be less convenient. In recent times, HRV can also be measured using interbeat interval (IBI) data obtained through methods like chest belts or Polar heart rate devices. However, these methods may introduce artifacts and lack the precision of ECG-based HRV analysis. Photoplethysmography (PPG) is an emerging technique that measures variations in absorbed light caused by arterial blood flow and can be used as pulse rate variability (PRV). PRV has shown a strong correlation with HRV and can be measured quickly and easily using smartphones or low-cost devices.Figure 2.8 shows synchronized (ECG) and (PPG) waveforms. PPG has gained popularity due to advances in optoelectronics and digital signal processing, making it a non-invasive, cost-effective, and user-friendly alternative to ECG-based HRV analysis. In clinical settings where

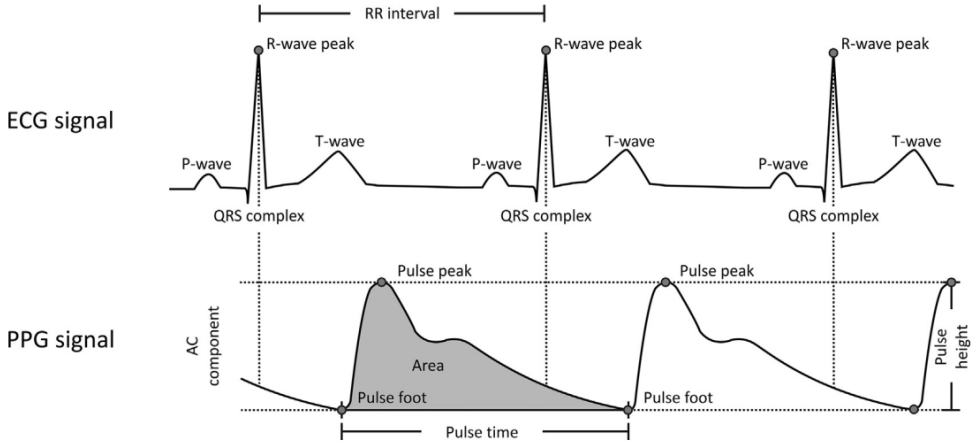



Figure 2.8: Typical synchronized electrocardiogram (ECG) and photoplethysmographic (PPG) waveforms and their respective components [22].

pulse oximeters are readily available, incorporating HRV analysis via PPG provides advantages over ECG, especially when metal-containing sensors are restricted. PPG sensors are typically attached to a finger or earlobe, requiring fewer leads and electrodes compared to ECG, making them a practical choice for various monitoring situations [28][22]. Figure 2.9 shows an Illustration of typical R-wave peak detection observed from ECG signals (A), corresponding heartbeats detected on PPG signals (B), and the resulting heartbeat intervals from both origins (C).

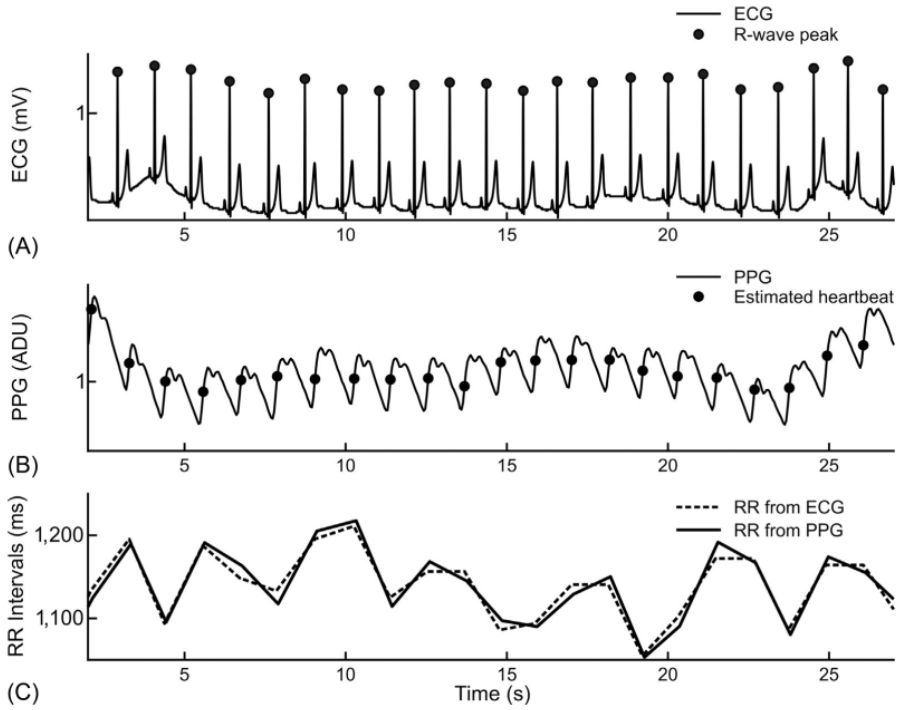



Figure 2.9: Illustration of typical R-wave peak detection (cardiac muscle contraction) observed from ECG signals (A), the corresponding heartbeats detected on PPG signals (B) and the resulting heartbeat intervals from both origins (C) [22].

HRV assesses beat-to-beat variability and provides time-domain, frequency-domain, and non-linear domain indices for analysis, measuring the variability in the Inter-Beat-Interval (IBI) between successive heartbeats. Table 2.2 describes briefly the commonly used HRV/PRV indices and their physiological origin. HRV observation periods range from over 1 minute to less than 24 hours [25].

Table 2.2: Summary of the common heart rate variability parameters and their physiological origin [29].

| Domain | Variable | Description | Physiological origin |
|---|---|---|---|
| Time | SDNN | Standard deviation of all R–R intervals | Cyclic components responsible for heart rate variability |
| | RMSSD | Root mean square of successive differences | Vagal tone |
| | pNN50 | Percentage of successive normal sinus RR intervals more than 50 ms | Vagal tone |
| | HR | Mean of Heart rate | - |
| | MeanNN | Mean of all R-R interval | - |
| Frequency-domain | LF | Low frequencies | Mix of sympathetic and vagal activity, baroreflex activity |
| | HF | High frequencies | Vagal tone |
| Non-linear indices | SD1 | Standard deviation – Poincaré plot Crosswise | Unclear, depicts quick and high frequent changes in heart rate variability |
| | SD2 | Standard deviation – Poincaré plot Lengthwise | Unclear, depicts long-term changes in heart rate variability |

### 2.4.4 EDA signal

Electrodermal activity (EDA) reflects sympathetic innervation of sweat glands and involves tonic and phasic changes in electrical conductance on the skin. Sudomotor nerves, part of the sympathetic nervous system, control eccrine sweat glands, influencing sweat production and duct opening. EDA is primarily linked to sympathetic nervous system activity and is associated with various functions, including gross movements, thermoregulation, emotional processes, and attention. Skin conductance, a measure of EDA, has been used extensively in psychological research to assess the skin's electrical resistance. It increases in response to stressors, and basal resistance decreases. Skin conductance reflects both immediate and long-term emotional arousal, making it a valuable indicator of autonomic nervous system activity[30][31]. Skin conductance responses (SCRs) consist of a tonic component (skin conductance level, SCL) and a phasic component (conductance responses). SCL fluctuates slowly, reflecting sweat diffusion, while SCRs result from rapid sweat release through duct openings, triggered by sympathetic nerve bursts, and it is often used as a general measure of psycho-physiological stress, while SCLs indicate overall sympathetic activity. Peak amplitude, measured as the difference between the SCR's peak and valley, is used in skin conductance response analysis. To enhance accuracy, a deconvolution analysis method has been applied to extract individual SCRs and measure peak amplitudes more precisely[32]. Figure 2.10 shows EDA data decomposition into tonic and phasic components and Figure 2.11 shows A typical skin conductance response (SCR) and illustration of some derived measures (Features).

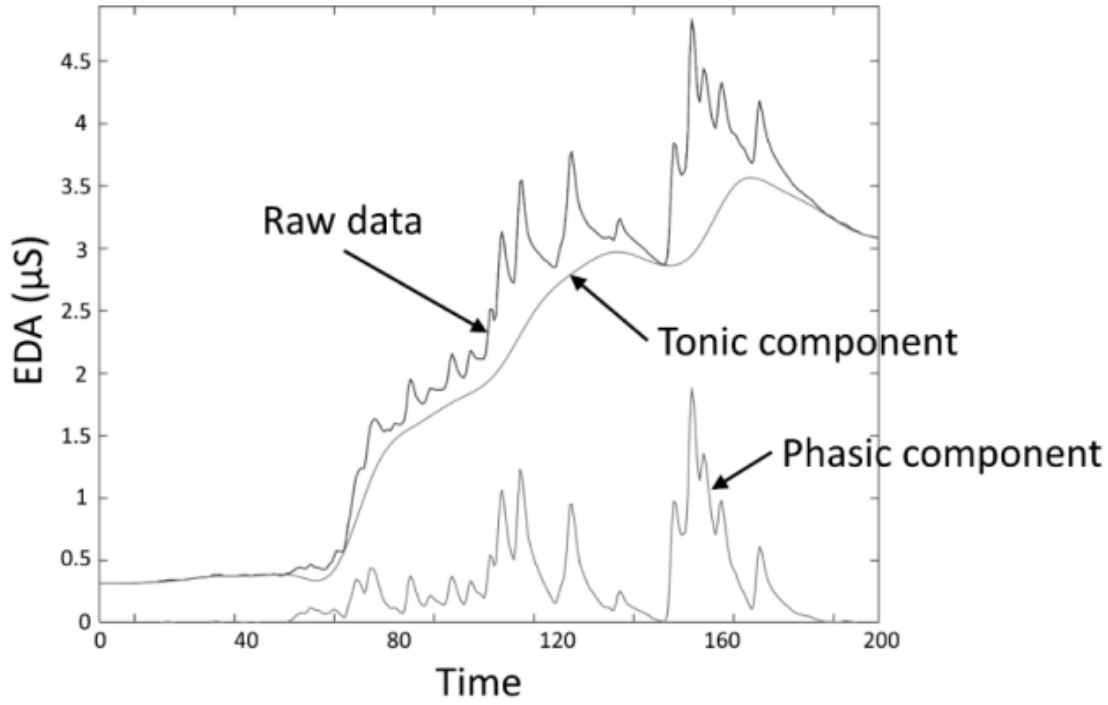



Figure 2.10: EDA data decomposition into tonic and phasic components [33].

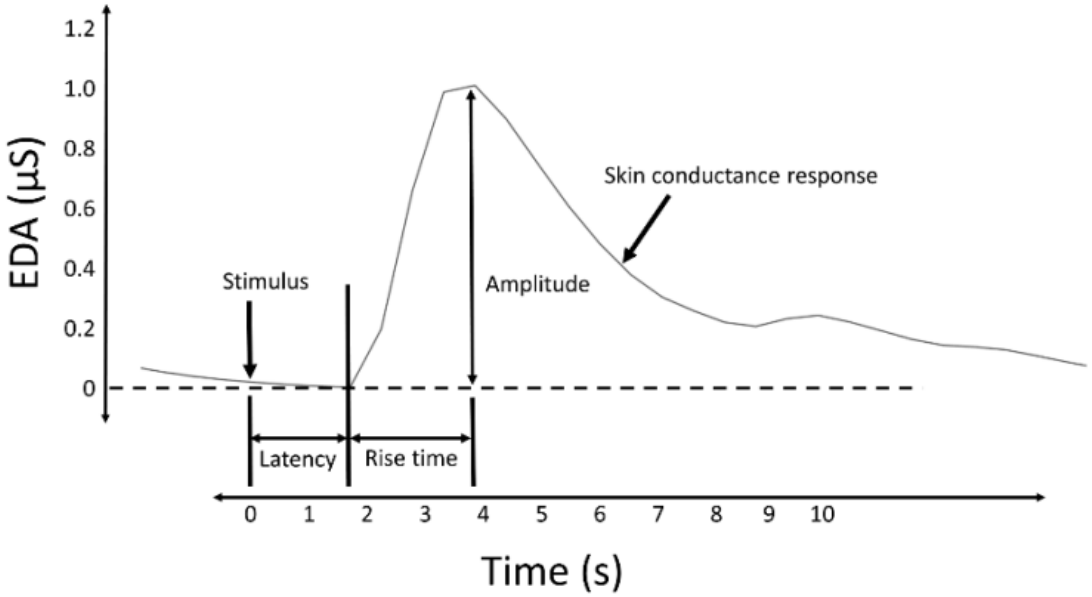



Figure 2.11: A typical skin conductance response (SCR) and illustration of some derived measures [33].

# Chapter 3

# Artificial intelligence

## 3.1 Introduction

Artificial intelligence can be defined as the capability of computer-controlled machines or robots to carry out tasks typically associated with intelligent beings. It involves software and hardware methods that mimic human behavior and thinking. Weak AI and strong AI, also known as general artificial intelligence, are two categories of AI based on the system's degree of intelligence compared to that of a human. Weak AI, or soft AI, is commonly used in practical applications and can efficiently solve specific problems with acceptable accuracy. Strong AI, or general artificial intelligence, is the focus of research. One of the major components of AI is machine learning (ML), which includes a subset of algorithms called deep learning (DL). DL is a relatively new addition to the ML family and is based on artificial neural networks. The interconnection between AI, ML, and DL is shown in Figure 3.1. Apart from ML, other significant subfields of AI include natural language processing, text and speech synthesis, computer vision, robotics, planning, and expert systems [34] [35]. These domains of AI are illustrated in Figure 3.2.

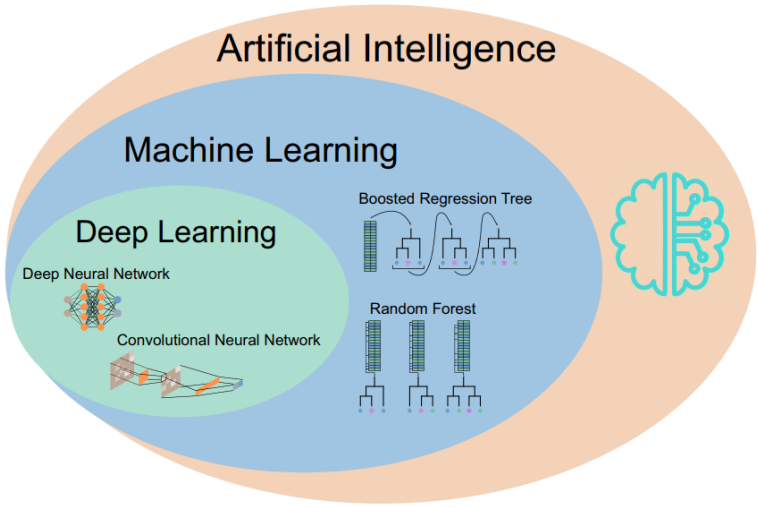



Figure 3.1: Relationship between artificial intelligence (AI), machine learning (ML), and deep learning (DL) [35].

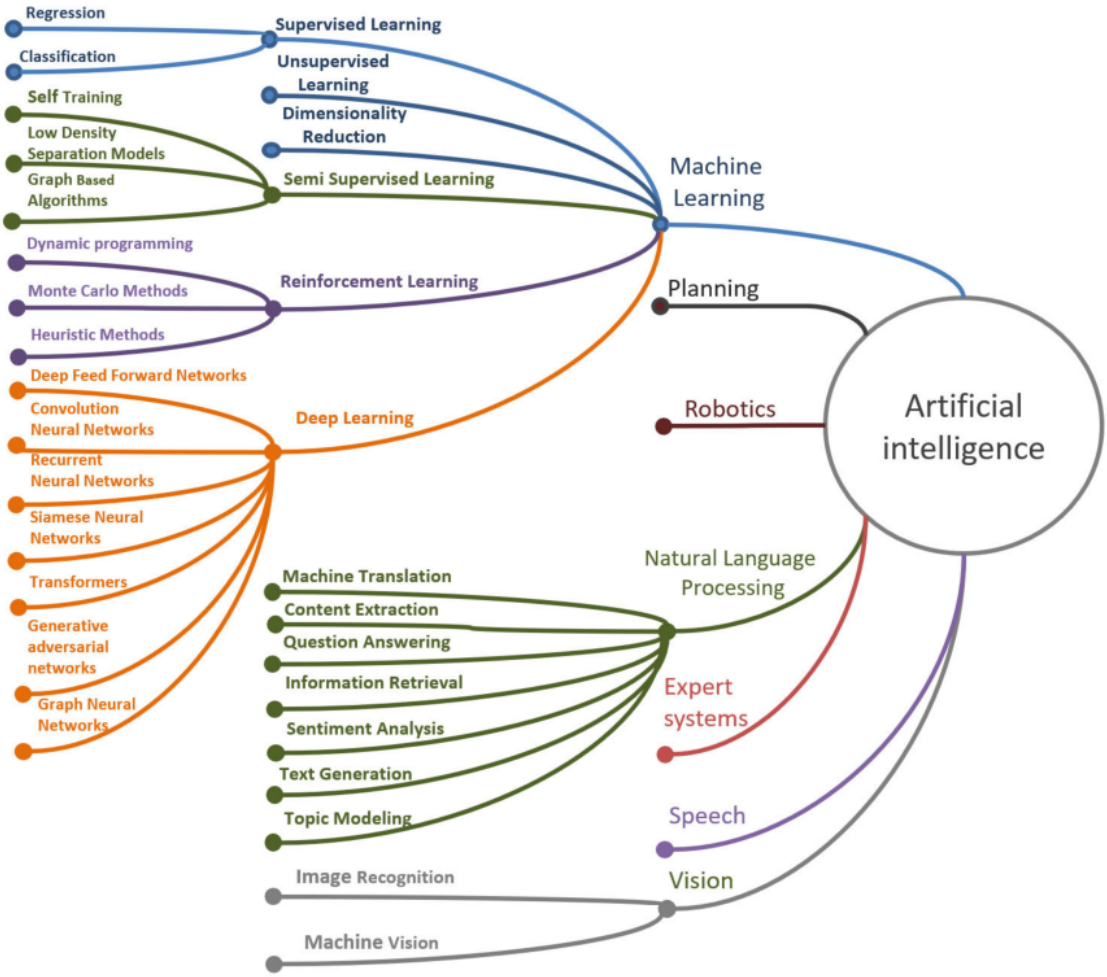



Figure 3.2: Subsections of artificial intelligence [34].

## 3.2 Features selection

A key idea in modeling is feature selection, which can improve a model's performance by eliminating unnecessary features. Feature selection becomes one of the crucial steps in building our stress detection model in order to reduce the complexity and the time needed for the execution of computations, which have been greatly increased due to the use of cross-validation. In order to enhance the effectiveness of stress detection, the most pertinent and significant features should be chosen. The ranking of feature importance was performed using two methods: Univariate feature ranking for classification using chi-square tests (Chi-test), in Matlab, and the Pearson's correlation coefficient with the Waikato Environment for Knowledge Analysis (WEKA) [36]. The former one is so-called because it is done on two distributions to determine the level of similarity of their respective variances. In its null hypothesis, it assumes that the given distributions are independent [37]. The Chi-square test can be written as

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{3.1}$$

where $\chi^2$ represents the calculated value of the chi-square test, $\sum$ denotes the sum, $O$ represents the observed number of events in each category, $E$ represents the expected number of events in each category, and $(O-E)^2$ represents the squared difference between the observed and expected number of events in each category. In our case, using the Matlab function *fscchi2*, which examines whether each predictor variable is independent of a response variable by using individual chi-square tests. A small p-value of the test statistic indicates that the corresponding predictor variable is dependent on the response variable, and, therefore is an important feature. We computed the predictor scores as –log($p$), with $p$ being the p-value. Therefore, a large score value indicates that the corresponding predictor is important. Then, we computed the mean value of the score and used it as a threshold.
Then, using WEKA, we applied a Pearson correlation coefficient to create rankings for each feature. Pearson's correlation coefficient is a measure of the linear relationship between two variables, X and Y. It ranges between -1 and +1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and +1 indicates a perfect positive linear relationship. The formula for the Pearson correlation coefficient is

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.2}$$

where $r_{xy}$ represents the Pearson correlation coefficient between X and Y, $\sum$ denotes the sum, $n$ is the sample size, $x_i$ and $y_i$ are the ith observations of X and Y, respectively, $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, respectively. In our case, the function *CorrelationAttributeEval* was applied to evaluate the worth of an attribute by measuring the correlation between it and the class. Any attributes with rankings below a cutoff of 0.10 were eliminated [38].

## 3.3 Machine Learning (ML)

Machine learning algorithms are utilized to analyze large, complex datasets and identify patterns by employing statistical, probabilistic, and optimization techniques. These algorithms find applications in various fields, such as text categorization, network intrusion detection, email filtering, credit card fraud detection, customer behavior analysis, manufacturing process optimization, and disease modeling. Supervised machine learning algorithms, which involve training a model using labeled data to predict outcomes for unlabeled examples, are predominantly used in these applications [39]. The main objective of machine learning (ML) is to develop predictive models that perform well on new data. A "good" model is one that can generalize and make accurate predictions beyond the data it was trained on. It is important to find the right balance between model complexity and flexibility to avoid overfitting, where the model becomes too specialized to the training data and fails to general-

ize. ML algorithms adjust their parameters to the training data while optimizing the bias-variance tradeoff. This tradeoff represents the relationship between model complexity and flexibility. The goal is to find the optimal level of complexity that allows the model to generalize well to new data. Figure 3.3 shows a Decision tree to assist in task identification. Given feature matrix X and a response vector y, the first decision is to choose between unsupervised (outcome y is unobserved) and supervised (outcome y is observed) learning. In the case of supervised learning, if y is discrete (e.g. species classes), it is a classification task, and if y is a continuous variable (e.g. biomass), it is a regression task. ML tasks are categorized into supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the algorithm is provided with examples of correct task execution, and the model is trained to minimize the differences between its actions and the correct actions. Common supervised tasks include classification (labeling data) and regression (predicting numerical variables). Unsupervised learning refers to tasks where no examples are provided, and the algorithms optimize a general loss function. Reinforcement learning involves training the ML algorithm by interacting with an environment, where learning depends on executed actions and their consequences. Various model classes and architectures can be used in ML. In supervised learning, neural networks, regression and classification trees, and distance-based methods are commonly used. In unsupervised learning, model classes include agglomerative hierarchical methods and those requiring a specified number of clusters.
Training a model in supervised and reinforcement learning involves two steps. The first step is defining a loss function that measures the algorithm's performance in solving a specific task. The loss function differs for classification and regression tasks. The second step is using an optimizer to update the algorithm's parameters and improve its performance. In unsupervised learning, similarities between observations are often used to determine grouping [35].

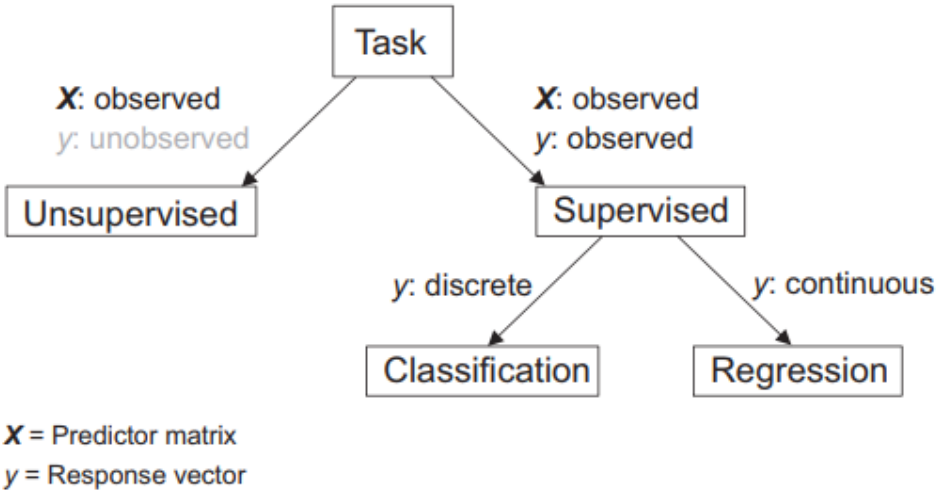



Figure 3.3: Decision tree to assist in task identification in ML [35].

## 3.4 Data splitting

Once the data has been collected, cleaned, investigated, and subjected to feature engineering, it is essential to establish an evaluation strategy for the models that do not rely on the training data. Two common approaches for this purpose are creating a separate "holdout" dataset or performing cross-validation. These strategies ensure that the model's performance is assessed on unseen data, enabling a more reliable evaluation.

### 3.4.1 Hold-Out method

Holdout validation is a common approach for evaluating machine learning models shown in Figure 3.4 The blue and green samples represent different samples from 2 different classes. In the holdout method, samples are randomly assigned to the training (purple box), validation (yellow box), or test (orange box) sets. When the dataset used for training and evaluation of the ML model is small, the performance measures using validation and test sets are sensitive to the composition of these sets, and the resulting performance measures often are not reliable. The proportions of data allocated to each set depend on factors such as the number of data points, data variability, and model characteristics. Typically, 70% of the data is used for training, 15% for validation, and the remaining 15% for testing, although these percentages can vary. The training and validation sets are used for model building. The training set is used to learn the model parameters, which are the properties or variables of the model that are adjusted during training. Examples of model parameters include weights and biases in a neural network or coefficients in a linear regression model. On the other hand, hyperparameters are not learned from the training set. They are determined using the validation set and include factors such as the number of layers in a neural network or the regularization values in a regression model. Hyperparameter tuning, the process of optimizing the hyperparameters, can be computationally demanding but can be done in a parallel and automated manner. After the model is trained and fine-tuned, it is evaluated on the test set to estimate the model's generalization error, which reflects its performance on unseen data. It is crucial that the test data are not used during training and fine-tuning to obtain an unbiased estimate of the model's performance. Holdout validation is commonly used for training deep learning models with large-scale datasets due to its computational efficiency. However, it has limitations when applied to small datasets. A small test set may not provide a reliable estimate of model performance, and the choice of the test set can affect the performance measures. Selecting a representative test set for small datasets can be challenging. Additionally, using a larger test set reduces the available samples for training, negatively impacting model performance. The choice of the validation set can also influence the model's ability to generalize when fine-tuning the model using holdout validation [40].

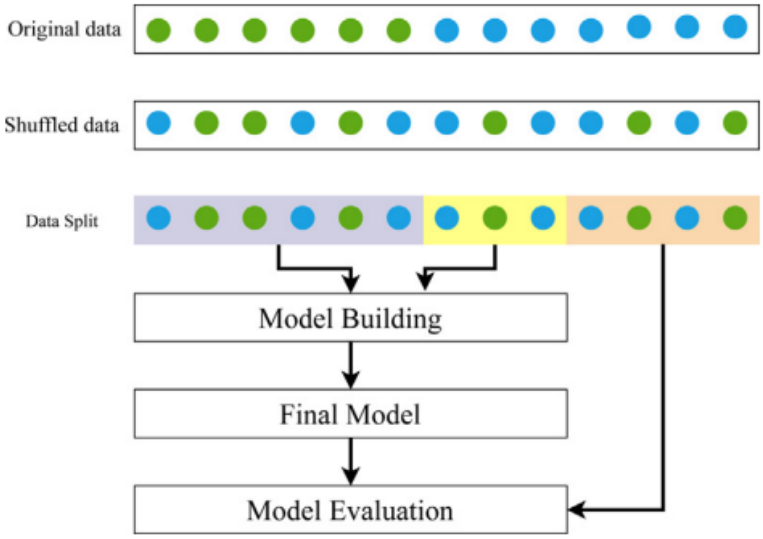



Figure 3.4: The holdout method [40].

### 3.4.2 Cross-validation (K-folds) method

When the dataset is limited in size, cross-validation techniques are often employed to address the associated limitations. In k-fold cross-validation, the value of k is specified, and the dataset is divided into k subsets or folds shown in Figure 3.5. One fold is used for testing, while the remaining k-1 folds are used for training the model. This process is repeated k-1 times, ensuring that each subset is used for testing. The choice of k determines the train-to-test ratio, with common values being 5 (80% training, 20% testing) or 10 (90% training, 10% testing). However, to determine the optimal value of k, it is recommended to perform repeated cross-validation and evaluate the model's performance across different values of k [41].

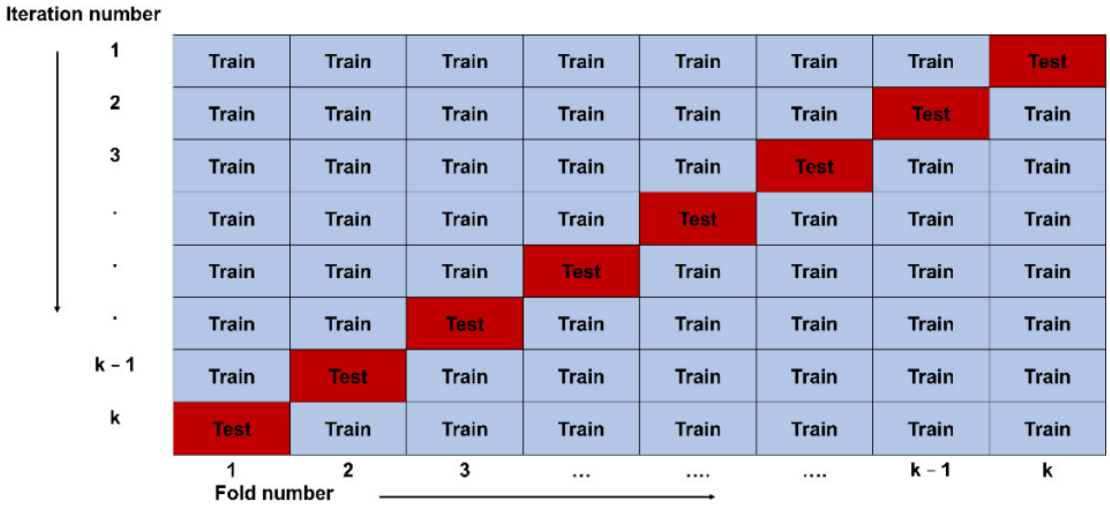



Figure 3.5: k-fold cross-validation method [41].

## 3.5 Machine Learning Classifiers

Supervised learning algorithms are well-suited for two main types of problems: classification and regression. In classification problems, the output variable is discrete and divided into different categories, such as colors or disease diagnoses. On the other hand, regression problems involve predicting a real-valued output variable, such as the risk of a specific health condition. In the upcoming sub-sections, a brief overview of commonly used supervised machine-learning classifiers for disease prediction will be provided. Figure3.6 illustrates an overall of classification process for Anxiety disorder.

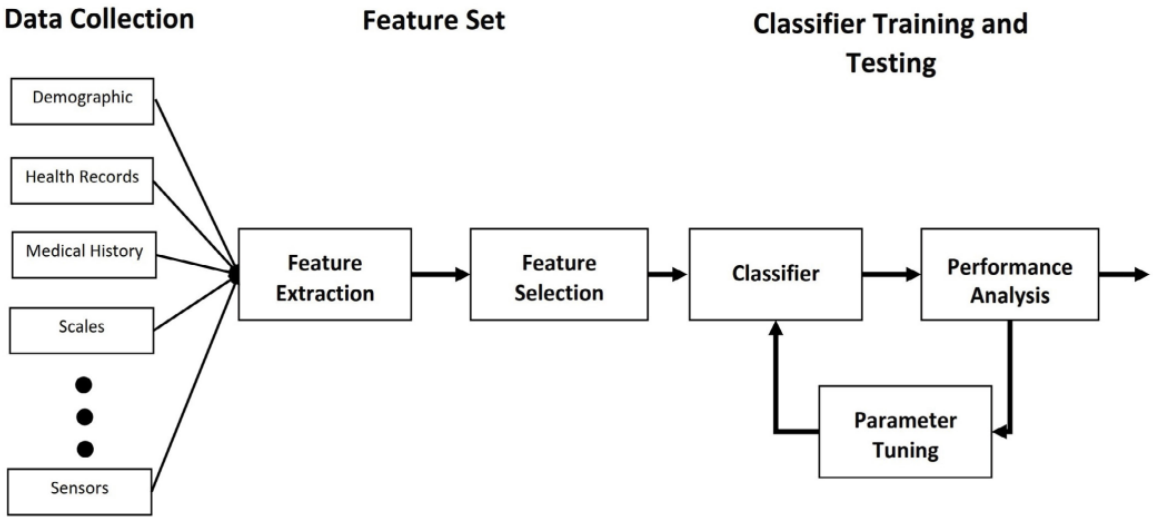



Figure 3.6: Classification process for Anxiety disorder [42].

### 3.5.1 Support vector machine (SVM)

The Support Vector Machine (SVM) algorithm is capable of classifying both linear and non-linear data. It accomplishes this by mapping each data item to an n-dimensional feature space, where n represents the number of features. The algorithm then identifies a hyperplane that effectively separates the data items into two classes while maximizing the margin between the classes and minimizing classification errors.
In more detail, each data point is represented as a point in the n-dimensional space, with each feature value corresponding to a specific coordinate. The objective of SVM is to find the hyperplane that best separates the two classes, maximizing the distance between the hyperplane and its nearest data points from each class. This hyperplane with a maximum margin serves as the decision boundary for classification. Figure 3.7 provides a simplified illustration of an SVM classifier [39].

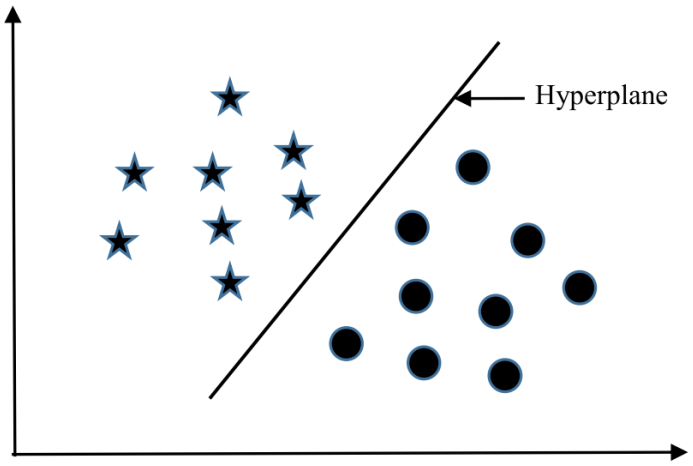



Figure 3.7: A simplified illustration of how the support vector machine works [39].

### 3.5.2 Logistic regression (LR)

Logistic regression (LR) Figure 3.8 is a widely used supervised classification method. It differs from linear regression as it utilizes a sigmoidal curve instead of a straight line. LR is an extension of ordinary regression and is specifically designed for modeling dichotomous variables, representing the presence or absence of an event. LR helps in estimating the probability of a new instance belonging to a specific class. The output of LR is a probability value ranging between 0 and 1. To use LR as a binary classifier, a threshold is chosen to distinguish between the two classes. For instance, if the probability value for an input instance exceeds 0.50, it is classified as "class A"; otherwise, it is classified as "class B". LR can also be extended to handle categorical variables with more than two values, known as multinomial logistic regression [39] [43].

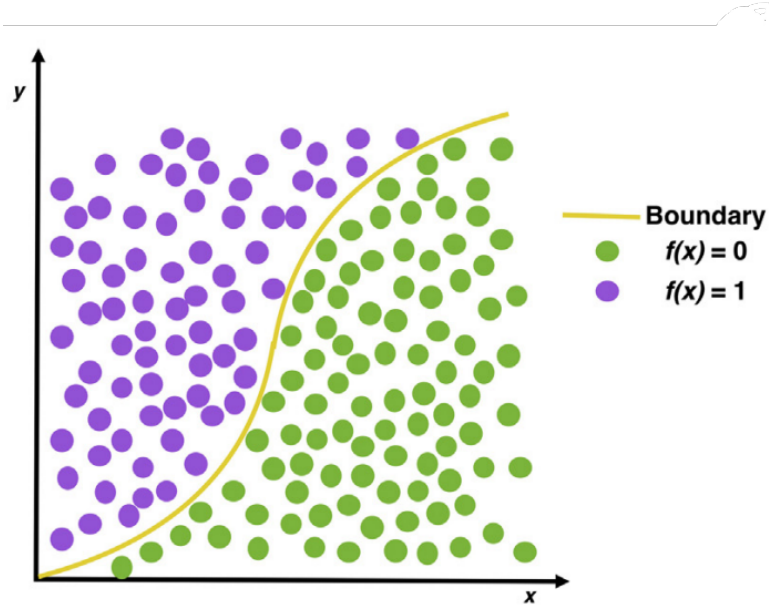



Figure 3.8: Graphical representation of logistic regression [43].

### 3.5.3 Decision tree (DT)

The decision tree (DT) algorithm is one of the earliest and widely used machine learning algorithms. It represents the decision-making process through a hierarchical tree structure, where each node corresponds to a test on input variables or attributes. The tree starts with a root node and branches based on the outcomes of the tests, leading to internal nodes and eventually reaching leaf nodes that represent decision outcomes.
Decision trees are known for their interpretability and ease of learning. They are commonly employed in medical diagnostic protocols due to their quick learning capabilities and straightforward interpretation. When classifying a sample using a decision tree, the outcomes of tests at each node along the path provide the necessary information to determine its class [39]. An example of a decision tree with its components and rules is illustrated in Figure 3.9.

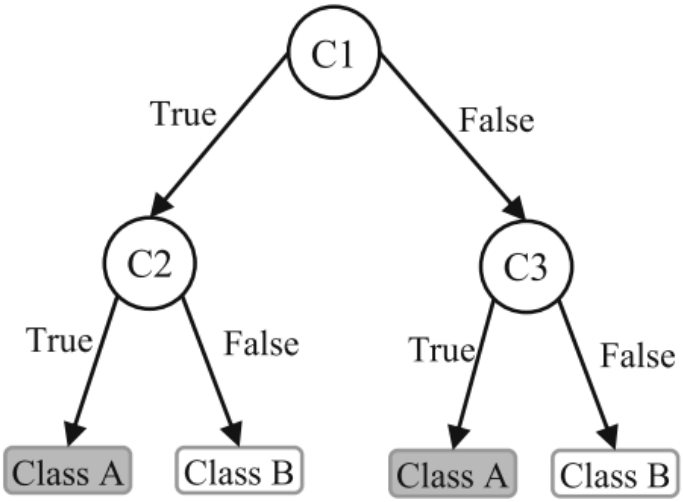



Figure 3.9: An illustration of a Decision tree [39].

### 3.5.4 Random forest (RF)

A random forest (RF) Figure 3.10 is an ensemble classifier that consists of multiple decision trees, similar to how a forest is a collection of many trees. Deep decision trees can often lead to overfitting, where they become too specific to the training data and result in high variation in classification outcomes for small changes in input data. They can also be sensitive to the training data, making them less reliable when applied to a test dataset. In an RF, the individual decision trees are trained on different subsets of the training dataset. When classifying a new sample, the input vector of that sample is passed through each decision tree in the forest. Each tree considers a different subset of the input vector and produces a classification outcome. The forest then combines these outcomes by either taking a majority vote (for discrete classification outcomes) or averaging the results (for numeric classification outcomes).

By considering the outcomes from multiple decision trees, the RF algorithm reduces the variance that can arise from relying on a single decision tree for the same dataset [39].

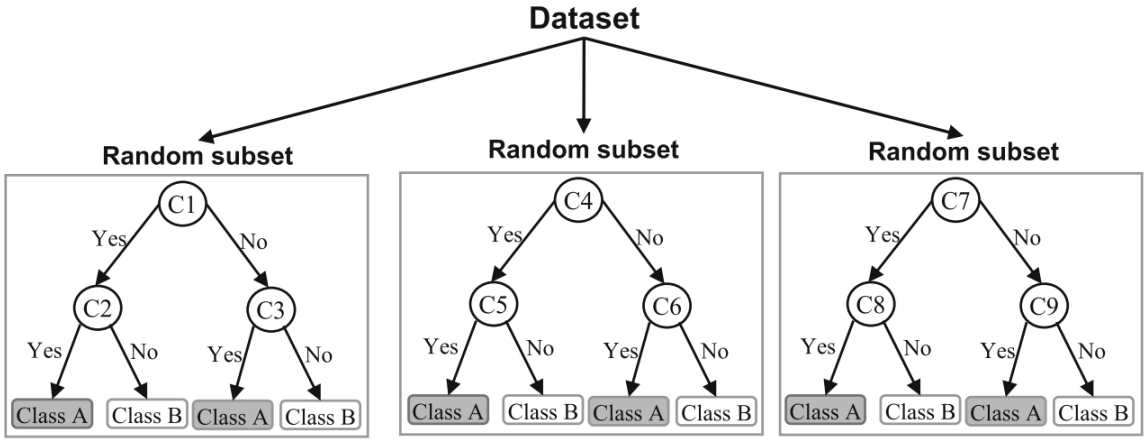



Figure 3.10: An illustration of a Random forest that consists of three different decision trees [39].

### 3.5.5 Naïve Bayes (NB)

Naïve Bayes (NB) is a classification technique based on Bayes' theorem, which describes the probability of an event based on prior knowledge of related conditions. The NB classifier assumes that features within a class are independent of each other, even though there may be interdependencies among the features of that class. The working principle of NB can be illustrated using an example in Figure 3.11 of classifying a new object (white circle) into either the 'green' or 'red' class. In this example, there are 40 'green' objects and 20 'red' objects, making it reasonable to believe that a new object is twice as likely to belong to the 'green' class based on the prior probabilities. To classify the 'white' object using NB, a circle is drawn around it, encompassing several points chosen prior, regardless of their class labels. In the example, four points (three 'red' and one 'green') are considered. The likelihood of the 'white' object given 'green' is calculated as 0.025, and the likelihood of 'white' given 'red' is 0.15. To obtain the final classification, the prior probabilities and likelihood values are combined using the 'multiplication' function, resulting in posterior probabilities. In the example, the posterior probability of 'white' being 'green' is 0.017, and the posterior probability of 'white' being 'red' is 0.049. Based on these probabilities, the 'white' object would be classified as a member of the 'red' class according to the NB technique.
Overall, NB utilizes prior probabilities, likelihood values, and Bayes' theorem to classify new objects by combining different sources of information [39].

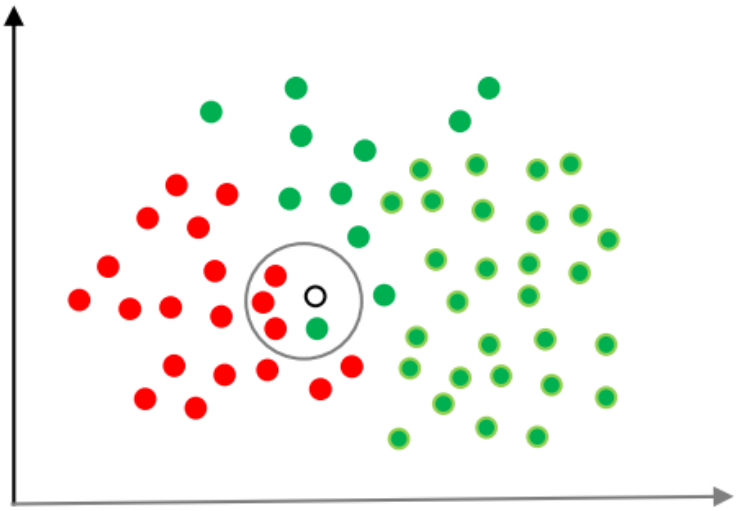


Figure 3.11: An illustration of the Naïve Bayes algorithm [39].

### 3.5.6 K-nearest neighbor (K-NN)

The K-nearest neighbor (KNN) algorithm is a simple and early classification algorithm. It can be considered a simpler version of the Naïve Bayes (NB) classifier as it does not involve probability values. In KNN, 'K' represents the number of nearest neighbors that are considered to determine the classification of a new object. The KNN algorithm does not rely on probabilistic calculations like NB. Instead, it identifies the K nearest neighbors based on a chosen distance metric and assigns the new object to the class that is most common among its neighbors. The value of 'K' can significantly impact the classification results for the same sample object. For instance, in Figure 3.12, the KNN algorithm is illustrated classifying a new object. When K = 3, the new object (star) is classified as 'black', while it is classified as 'red' when K = 5 [39].

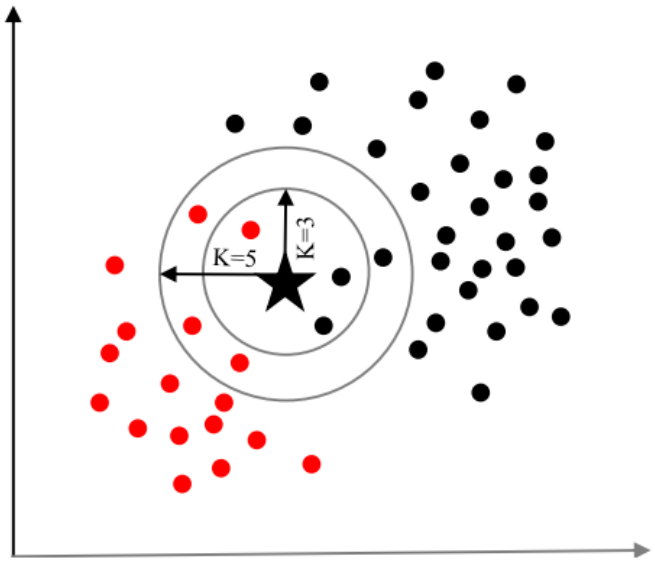



Figure 3.12: A simplified illustration of the K-nearest neighbor algorithm [39].

### 3.5.7 Artificial Neural Network (ANN)

Artificial neural networks (ANNs) are machine learning algorithms that draw inspiration from the functioning of the neural networks in the human brain. They were initially proposed and gained popularity in the 1980s. In the biological brain, neurons are interconnected through multiple axon junctions, forming a graph-like architecture. These connections can be rewired, allowing for adaptation, information processing, and storage. Similarly, ANN algorithms can be represented as interconnected nodes. The output of one node serves as input to another node for further processing, following the interconnections. Nodes are organized into layers, including input, output, and potentially one or more hidden layers. Nodes and edges in ANNs have weights that determine the strength of communication between them. These weights can be adjusted during training, either amplifying or weakening the signal strengths. Through repeated training, ANNs learn to make predictions for test data by adapting the matrices, node values, and edge weights. Figure 3.13 illustrates an ANN with two hidden layers, showcasing its interconnected nodes [39].

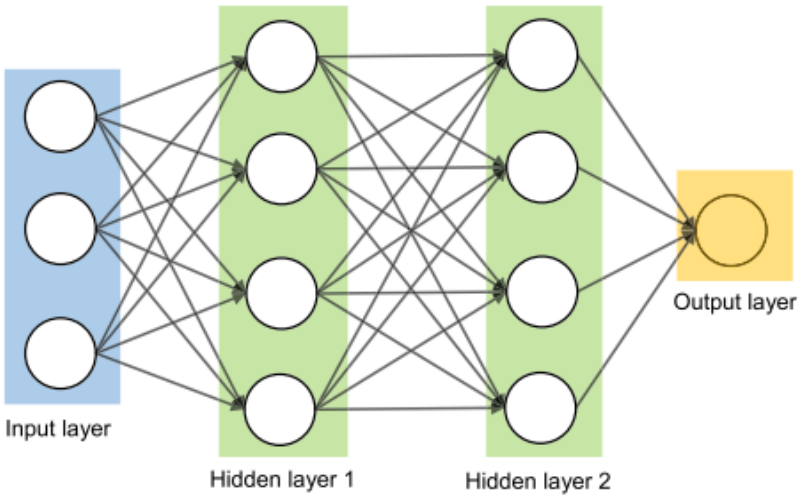



Figure 3.13: An illustration of the artificial neural network structure with two hidden layers [39].

## 3.6 Deep Learning

Deep learning, which has gained prominence since 2006, is a data processing method involving intricate multi-layer structures and nonlinear transformations. Notably, it has made significant breakthroughs in computer vision, speech recognition, and various other fields, earning recognition as a top technological advancement. Deep learning, inspired by the human neural network, abstracts data through successive layers, enabling tasks like target detection, classification, and segmentation. A key advantage is its ability to automate feature extraction, replacing manual efforts with

unsupervised or semi-supervised learning and hierarchical feature extraction algorithms [44]. In the context of medical data, especially medical imaging, deep learning, specifically Convolutional Neural Networks (CNNs), plays a pivotal role. Despite the diverse and fluctuating nature of medical data, CNNs excel in classification tasks, providing robust solutions. Various CNN architectures, such as AlexNet, VGGNet, and GoogLeNet, have demonstrated superior performance in image classification tasks. These networks employ techniques like convolution, pooling, and nonlinear activations (e.g., ReLU) to learn complex patterns efficiently [44][45]. Figure 3.14 illustrates General architecture of neural network and deep learning.

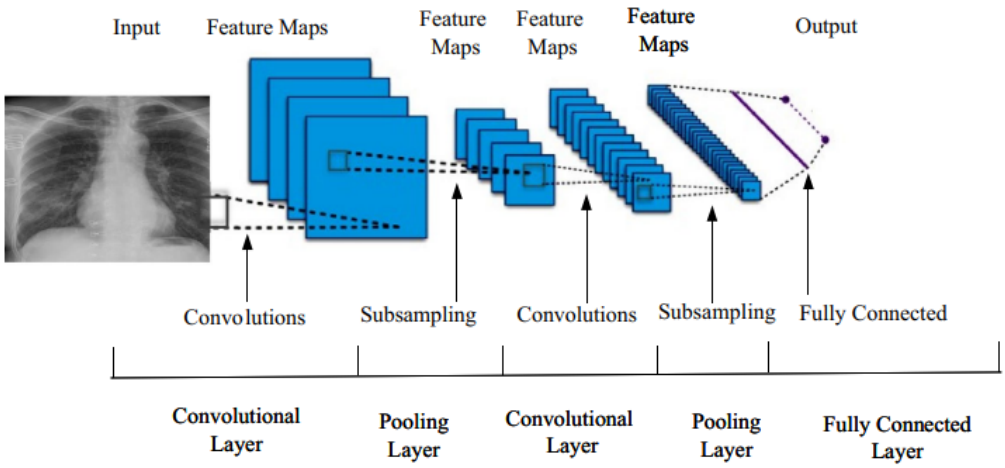



Figure 3.14: General architecture of neural network and deep learning [46].

Transfer learning is a rapid approach to creating precise models. Collecting an extensive dataset for training an entire CNN from the ground up is typically a challenge. Consequently, transfer learning emerges as the preferred option, where a pre-trained network, originally trained on a substantial benchmark dataset like ImageNet, is employed to address diverse problems, significantly reducing computational requirements. This technique involves transferring features learned by a primary (pre-trained) network to a target network, which is then fine-tuned using a specific target dataset, an example of pre-trained DNN: GoogLeNet, ResNet18, ResNet50, SqueezeNet, DenseNet-121 deep neural networks [47][46].

## 3.7 Hyperparameter Tuning

Hyperparameters are static model variables that control the model's behavior and architecture, set by the user before training. The most widely used techniques for hyperparameter tuning are Grid search, Random grid search, and Bayesian Optimization. In Grid Search, hyperparameters values are systematically predefined. In Random Grid Search, the hyperparameter space is randomly populated. In Bayesian Optimization, the chosen hyperparameters values are progressively optimized to

approximate a minimum [48]. Proper hyperparameter tuning is crucial for achieving better results with any machine learning model. In Figure 3.15 Illustration of how hyperparameter space (over two hyperparameters) is populated by different search schemes. Hyperparameter optimization significantly impacts the output of a machine-learning model, making it a critical step in its development. Without efficient hyperparameter optimization, individuals may randomly select hyperparameters and repeatedly train and evaluate the model, leading to a wasteful and inefficient process that consumes valuable time and resources [49]. This is where OPTUNA comes into play, as it automates the hyperparameter optimization process. Optuna is a versatile hyperparameter optimization framework that introduces the concept of "define-by-run" to dynamically create search spaces. It efficiently handles both independent and relational sampling methods, making it adaptable to various tasks and environments. Optuna also incorporates an efficient pruning mechanism to eliminate unpromising trial runs, ensuring cost-effectiveness. Its user-friendly setup and memory data structure storage make it a valuable tool for hyperparameter tuning, offering a powerful solution for optimizing machine learning models [50][49].

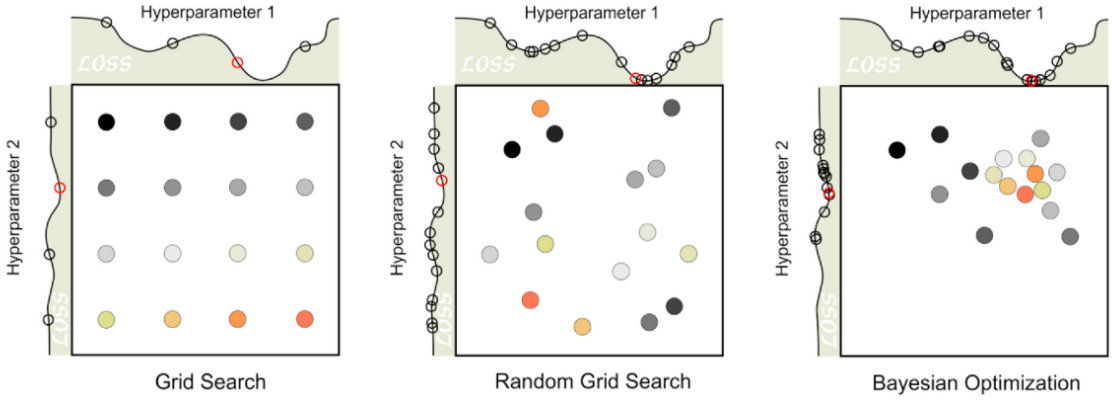



Figure 3.15: Illustration of how hyperparameter space (over two hyperparameters) is populated by different search schemes [48].

## 3.8 Explainable AI

In various applications, the need to understand why a machine learning model makes specific predictions is becoming increasingly important, especially as complex black-box models deliver high accuracy but are challenging to interpret. This challenge is particularly critical in fields like healthcare, where ML models are used for early disease detection. To address this interpretability issue, Explainable Artificial Intelligence (XAI) tools like Lime, Dalex, and SHAP have emerged. In this study, SHAP (SHapley Additive exPlanations) is employed, which is considered a state-of-the-art XAI method. SHAP is rooted in game theory and assigns unified importance scores to features in ML models. It helps in understanding how different features impact

model predictions, either globally across all classes or locally for individual observations, thus enhancing transparency compared to traditional feature importance techniques. This approach aids in demystifying complex model behavior, making it invaluable for applications like healthcare. Game theory involves two essential components: a game and its participants. In the context of a classification model, the "game" represents the process of generating the model's outcomes. In this analogy, the "players" take on the role of the features within our model. Shapley analysis assesses and quantifies the contribution of each "player" to the overall "game," while SHAP analysis precisely measures the impact of each feature on the model's predictions [51][52].

## 3.9 Classifier performance index

The diagnostic ability of classifiers is commonly evaluated using the confusion matrix and the receiver operating characteristic (ROC) curve. The confusion matrix, also known as the error or contingency matrix, provides a framework for assessing classifier performance Figure 3.16 (a). It includes measures such as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These measures help determine the accuracy of the classifier's predictions. The most commonly used to analyze the performance of classifiers, including those that are based on supervised machine learning algorithms are metrics of Accuracy (3.3), Precision (3.4), Recall (3.5), and F-measure (3.6).
The ROC curve is a fundamental tool for evaluating diagnostic tests and is created by plotting the true positive rate against the false positive rate at different threshold settings. The area under the ROC curve (AUC) is a commonly used metric to assess the predictability of a classifier. A higher AUC value indicates a superior classifier, while a lower value suggests lower predictive accuracy. The AUC value is determined by the coverage area under the ROC curve Figure 3.16 (b).
In addition to the confusion matrix and ROC curve, other measures such as the running mean square error (RMSE) and mean absolute error (MAE) are used to evaluate classifier performance. RMSE represents the mean value of squared errors between actual and predicted values, while MAE indicates the absolute value of the difference between actual and predicted values. These measures provide further insights into the accuracy of the classifiers [39].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.3}$$

$$Prec = \frac{TP}{TP + FP} \tag{3.4}$$

$$Rec = \frac{TP}{TP + FN} \tag{3.5}$$

$$F1 = \frac{2 * Prec * Rec}{Prec + Rec} = \frac{2 * TP}{2 * TP + FP + FN} \tag{3.6}$$

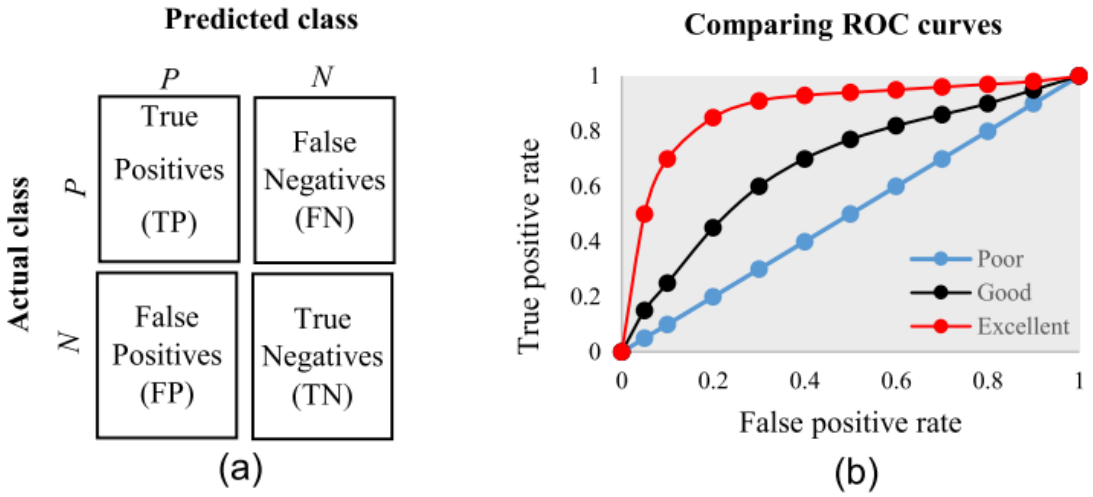



Figure 3.16: (a) The basic framework of the confusion matrix; and (b) A presentation of the ROC curve [39].

# Chapter 4

# Literature Review

## 4.1 Introduction

Stress is a common experience in modern work environments, and it can have negative effects on employee well-being and productivity. As a result, there has been increasing interest in developing tools and methods for detecting and managing stress in the workplace. Wearable devices and machine learning algorithms have been used in previous studies to detect stress, but these methods have limitations in terms of invasiveness, noise, and reliability in various environments. Therefore, this chapter aims to review recent research on stress detection methods in the workplace, focusing on approaches that acquire data from wearable sensors to improve accuracy and reduce limitations.

## 4.2 Method

To conduct this literature review, a comprehensive search was performed in electronic databases such as PubMed, Scopus, and Google Scholar. The search terms included "stress detection", "workplace", "wearable devices", "Heart rate variability", "Electrodermal Activity", "machine learning", and related keywords. Inclusion criteria were studies published in English language journals or conference proceedings from 2019 to 2023 that focused on stress detection using wearable devices in the workplace. Exclusion criteria were studies that did not meet the inclusion criteria or were duplicates. The screening of titles and abstracts was followed by full-text screening, and relevant data were extracted from the selected studies with a total number of fourteen. The extracted data included the Device name, signals used in stress detection, the used classifiers, type of stressors used, accuracy, and the number of subjects. The results of the literature review will be presented as a narrative synthesis and summarized in the Table 4.1.

## 4.3 Results

### 4.3.1 Rescio et al. (2023)

The study [53] conducted by Rescio et al. in 2023 aimed at designing an automated stress detection platform to address the new problems that arise owing to the human-machine interaction in Industry 4.0, by combining data from a wearable device and an environmental system. Twenty subjects wore a Shimmer device shown in Figure 4.1 designed to be minimally invasive with good signal stability and low noise, and a commercial camera (Logitech C920 HD Pro webcam) was added to improve the performance of the system. The protocol shown in Figure 4.2 consisted of several stressors, such as the Trier Social Stress Test, Mental Arithmetic Stress Test, Stroop color, Math, and Memory tests. Features were extracted from the analyzed PPG and GSR signals, as well as ambient features, using the camera. Subsequently, several supervised (Decision Trees (DT), Random Forest (RF), K-Nearest Neighbors(KNN)), unsupervised (K-means, Gaussian Mixture Model (GMM), Self-Organizing Map (SOM)) classifiers were used to evaluate the proposed system. Gaussian Mixture Model achieved the best performance with an accuracy of 77.4% for one level of stress and 75.1% for two levels of stress using the unsupervised approach, while Random Forest Model achieved the best performance with an accuracy of 94.9% for one level of stress and 91% for two levels of stress using the supervised approach.

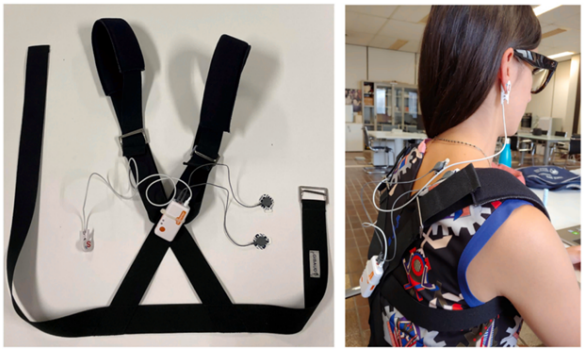


Figure 4.1: Prototype of wearable smart system of Rescio et al. study.

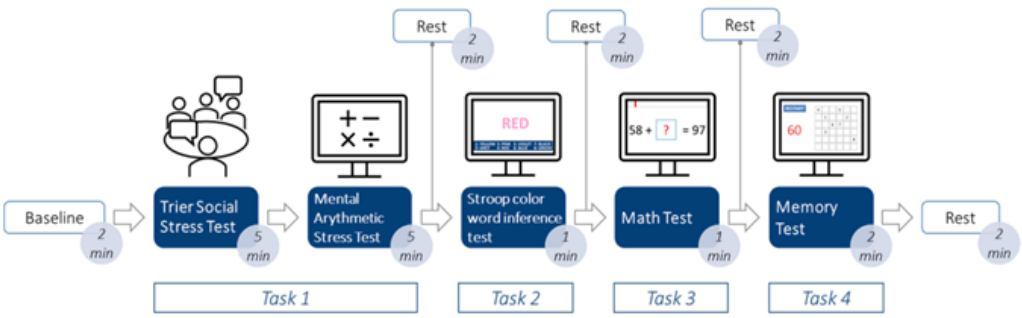



Figure 4.2: Stress-inducing protocol of Rescio et al. study.

### 4.3.2 Barki et al. (2023)

The study [54] carried out by Barki et al. in 2023 aimed at designing an ear-mounted photoplethysmography (PPG) system to detect mental stress. Fourteen participants wore the proposed device which mainly uses a MAX30102 pulse oximeter to obtain a PPG signal and a BNO055 accelerometer and gyroscope which was used to remove motion artifacts from the PPG. Figure 4.3 shows the prototype of the proposed system. Stress was induced using the Stroop color-word test and mental arithmetic tasks. After analyzing the PPG signal, it was transformed into scalograms using continuous wavelet transform (CWT). Subsequently, stress classification was performed using a convolutional neural network. The results, in terms of accuracy, were 92.04%; and 96.02% after adding white Gaussian noise to the raw PPG signals.

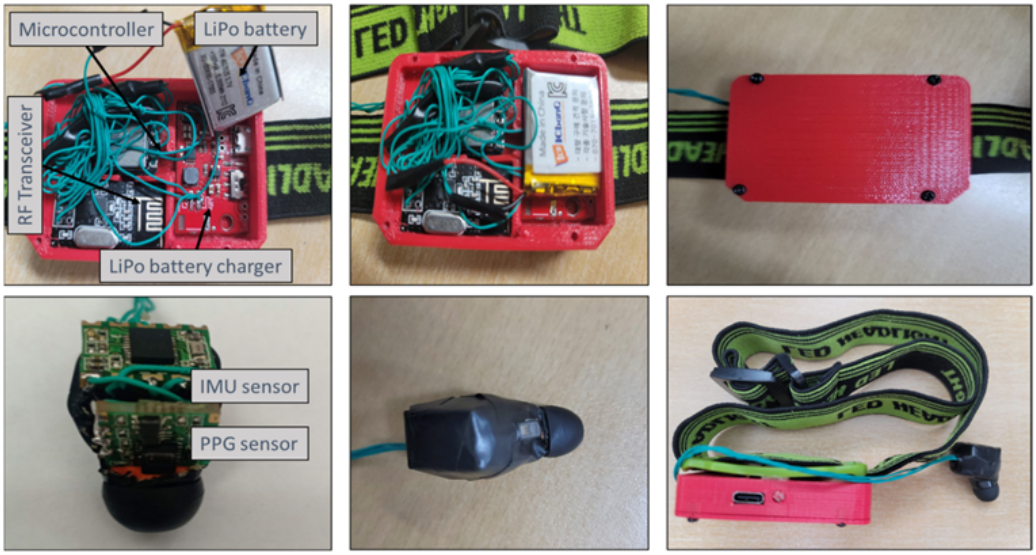



Figure 4.3: Prototype of the proposed system of Barki et al. study.

### 4.3.3 Mach et al. (2022)

In the study [55] carried out by Mach et al. in 2022, a laboratory experiment consisting of an arithmetic task which is counting down or up steadily, and physical activity (sitting vs. stepping) with 52 participants was conducted. The aim of this study was to assess mental workload via heart rate measurement using the Samsung Gear S3 smartwatch, furthermore, to confirm these results a chest strap (1-channel ECG) was used. They found that the mean heart rate increased when participants performed the arithmetic task compared to the conditions with no arithmetic task during both conditions (sitting and stepping). However, during stepping, the congruency of the heart rate values conducted by the smartwatch and ECG chest strap was weak.

### 4.3.4 Seo et al. (2022)

In the [56] study done by Seo et al. in 2022, 24 participants wore a Zephyr chest strap equipped with a BioHarness module to acquire ECG and Respiratory signals. Furthermore, the subjects were sitting in front of a laptop and faced a camcorder screen in order to register facial information. The experiment lasted for 45 min and comprised two stages: an initial setting stage, and an actual experiment stage which is the Stroop task. The actual experiment consists of Relax, Easy Stroop, Recovery, Hard Stroop, and Recovery, 5 min for each, Figure 4.4 shows the experimental setup. Afterward, signal and image processing was done followed by a Deep Neural Network (DNN) classifier. The accuracy for two or three levels of stress classification was 73.3%, and 54.4% respectively.

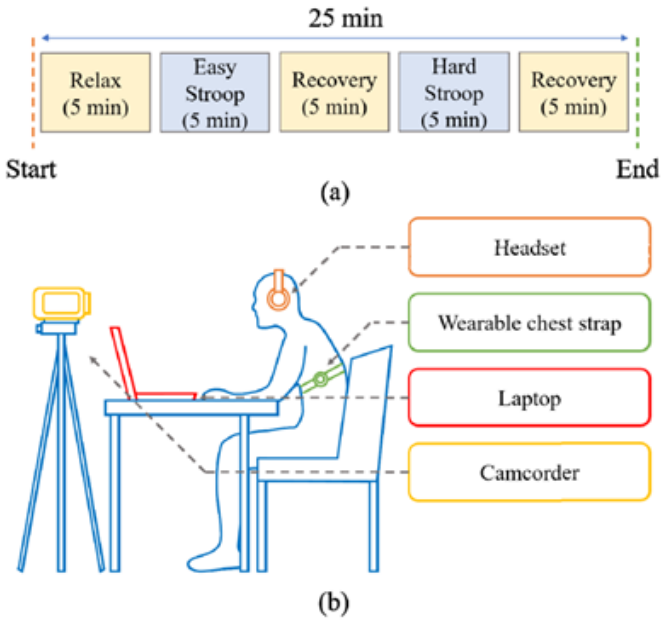



Figure 4.4: Experimental Protocol carried out by Seo et al.

### 4.3.5 Umer (2022)

The study [57] carried out by Umer in 2022, eight participants took part in the experiment that aimed to monitor physical and mental stress in the construction industry. Equivital EQ02 Life monitor was used to monitor physiological parameters (ECG, skin temperature, breathing, skin conductance). Figure 4.5 shows the placement of the sensors. The experiment was divided into two days: the first induce physical workload by manual handling of a 15-Kg backpack for 25 min, and the second induce mental workload using a digits-transformation task by hearing randomized four-digit number, for example, 7468 and replying to the number by adding one to each digit, for example, 8579 against 7468. The task continued with randomized numbers for 25 min. After that, the participants were asked to perform manual

material handling as mentioned before to induce mental and physical workload. The classification of physical and mental stress was performed simultaneously using machine learning algorithms and the best accuracy achieved using the bagged trees algorithm was 94.7% to predict physical and mental stress.

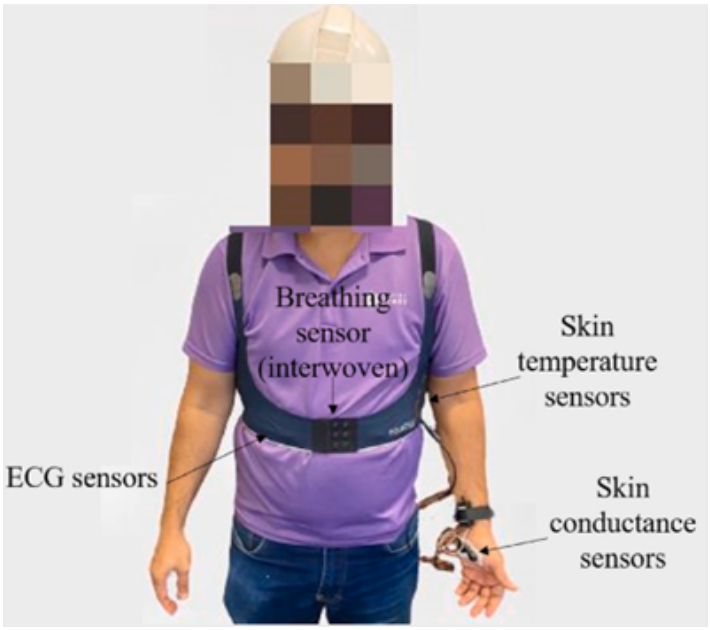



Figure 4.5: Sensors placement of Umer study.

### 4.3.6 Chalabianloo et al. (2022)

In [58] a study by Chalabianloo et al. in 2022, thirty-two subjects were subjected to a laboratory experiment that consisted of baseline, stress, recovery, and cycling sessions. Stress sessions were performed using the Stroop task. ECG and HR signals were recorded using seven different wearable devices simultaneously which are: BITalino (r)evolution board, Firstbeat Bodyguard2, Polar H10, Zephyr HxM, Empatica E4, Samsung Gear S2, and CoreSense, Figure 4.6 shows the placement of sensors. Support Vector Machine, Random Forest, Extremely Randomized Tree, and Light Gradient Boosting Machine were used to classify four classes: Baseline, Stress, Recovery, and Cycling. The best accuracies across most of the devices were obtained using an Extremely Randomized Tree classifier, for example, 88.26% for the BITalino device. Furthermore, in order to study the effects of multi-modality, the EDA signal was introduced using Empatica E4. After that, the same classifiers mentioned above were used. The accuracy obtained considering only HR was 83.89% using the Random Forest classifier, while when considering HR and EDA the accuracy became 90.62% using the Extremely Randomized Tree classifier.

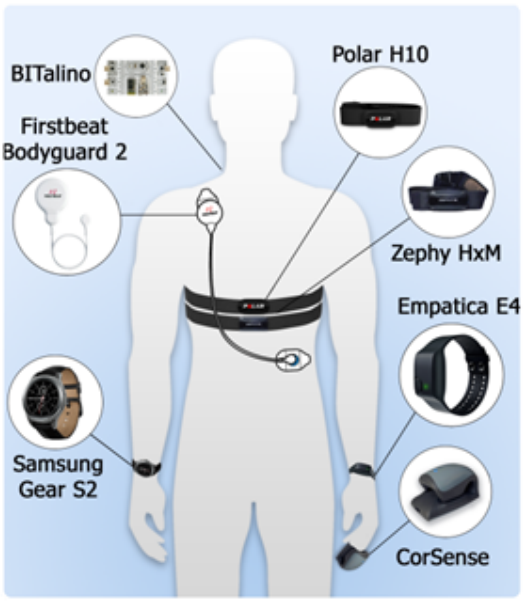



Figure 4.6: Sensors placement of Chalabianloo et. al study.

### 4.3.7 Li et al. (2022)

The study [59] by Li et al. in 2022 aimed to measure the workplace stress of nurses using Heart Rate Variability (HRV) analysis based on data derived from wearable ECG heart rate monitors. Seventeen nurses participated in the study and wore wireless heart rate monitor (myBeat-WHS-1, Union Tool Co., Ltd., Japan) to obtain ECG measurements during work time. After statistical analysis of HRV features, they found that Low-Frequency components (LF%) at work phase was significantly higher than at rest phase. In contrast, the natural logarithm of High-Frequency components (LnHF), and the squared root of the mean squared differences of successive NN intervals (RMSSD) at work phase were significantly lower than at rest phase. The results demonstrate the ability of stress detection using wearable sensors and HRV analysis.

### 4.3.8 Fauzi et al. (2021)

The study [60] done by Fauzi et al. in 2021 aimed at continuous stress detection of hospital staff using Empatica E4 smartwatch. WESAD dataset was considered in this study, which includes data from fifteen people. Features were extracted from EDA, Skin temperature, Acceleration, and Blood Volume Pulse signals. Furthermore, several machine learning classifiers were used such as Naïve Bayes (NB), Support Vector Machine (SVM), Neural Network (NN), K-Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT) in addition to ensemble methods which are technique trains numerous classification methods and then combine them using a particular approach. The best accuracy obtained using an individual classifier was 86.61% using RF classifier while for the ensemble technique

was 87.10% using the combination of RF, LR, and NN classifiers.

### 4.3.9 Dai et al. (2021)

In the study [61] carried out by Dai et al. in 2021, thirty-two subjects participated in 2 hours of laboratory and 24 hours of field-based experiments. The aim of the study was comparing between objective and subjective stress detection models. First of all the subjects wore a Fossil Gen4 Explorist smartwatch which is equipped with a photoplethysmogram (PPG) sensor as well as a six-axis inertial measurement unit (IMU). The Laboratory experiment included several stages such as resting, speech, recovery, math, and cold stressors. Moreover, the subjects were asked to continue wearing the smartwatch to complete 24 hours field-based experiment and collect data regarding stress in free-living situations. Support Vector Machine (SVM), Random Forest (RF), AdaBoost, gradient boosting (GB), and logistic regression classifiers were used to detect stressed or non-stressed periods in both objective and subjective stress models. The best accuracies obtained were: 82.6% for the objective stress model using the SVM classifier and 79.8% for the subjective model using the RF classifier.

### 4.3.10 A S et al. (2020)

In [62] a study by A S et al. in 2020 aimed to stress detection during the pre-surgery period based on Electrodermal activity using wrist wearables. Forty-one subjects wore ADI-VSM wrist-watch, which enables continuous monitoring of Electrodermal activity. The physiological data were collected for an approximate duration of 3 hours prior to the scheduled surgery. After signal processing and features extraction, an 85.06% of accuracy was obtained using a K-Nearest Neighbor classifier for three classes of stress (Low, Moderate, and High).

### 4.3.11 Said can et al. (2020)

In the study [63] carried out by said can et al. in 2020. Blood Volume Pressure, Skin Temperature, Electro Dermal Activity, IBI (Inter-beat Interval), and 3D Acceleration data were collected from 16 participants using Emaptica E4 smart band during daily activities interspersed with relaxation sessions like doing yoga, or mindfulness. After that, several classification algorithms were used in order to assess stress levels such as MultiLayer Perceptron, Random Forest, Linear Discriminant Analysis, Principal Component Analysis, and K-nearest Neighbors. The best accuracy achieved considering HRV, EDA, and accelerometer signals was 85.36% using the LDA algorithm for three classes (high stress, mild stress, and relax), while the best accuracy achieved by neglecting the relax class was 98% using MLP and RF algorithms.

### 4.3.12 Kaczor et al. (2020)

The study [64] carried out by Kaczor et al. in 2020 aimed at the objective measurement of physician stress in the emergency department using Empatica E4 smartwatch. Electro Dermal Activity, Acceleration, and heart rate signals were acquired from eight participants during clinical shifts (typically 8-10 hours). After that several machine-learning classifiers were used which are: decision trees, discriminant analysis, logistic regression, naïve Bayes, support vector machines, nearest neighbor, and ensemble classifiers, and the best accuracy obtained was 70% to detect stress during the working shift with respect to the baseline condition.

### 4.3.13 Kyriakou et al. (2019)

The work [65] by Kyriakou et al. in 2019 aimed to bridge the gap between laboratory settings and real-world field studies by introducing a new algorithm to detect moments of stress (MOS) using wearable physiological sensors. Eleven subjects wore an Empatica E4 device and were subjected to a laboratory experiment, an auditory stimulus was used to induce stress. The algorithm utilized GSR and ST signals to assess stress levels. Furthermore, in order to validate the algorithm, a real-world urban experiment was introduced. An accuracy of 84% was obtained using the proposed algorithm.

### 4.3.14 Suni Lopez et al. (2019)

In the study [66] done by Suni Lopez et al. in 2019, a laboratory experiment was conducted in order to detect stress in the office workplace, the experiment consisted of interacting with a laptop where the Stroop task was installed. Twelve subjects participated and were asked to wear the E4 smartwatch to collect EDA data, and headphones to interact with the environmental trigger (fire alarm). After signal filtering, aggregation, and discretization, an accuracy of 79.17% was obtained using statistical method classification.

hence, based on the literature review, the state-of-the-art and future studies to assess and detect stress levels by means of improving the analysis of signals during different situations, particularly in different work conditions such as speech-induced stress, physical-induced stress, etc. using several sensors such as Empatica E4 wrist-watch which was the most used device since it is less obtrusive and suitable for daily life and work stress assessment.

Table 4.1: Summary of Literature Review

| Study | Device | Signals | Method | Stressor | Accuracy | N. subjects |
|---|---|---|---|---|---|---|
| Rescio et al. (2023)[53] | Shimmer webcam | PPG, GSR, Ambient | DT, RF, K-NN, K-Means, GMM, SOM | Math, Speech, Stroop, Memory | 77.4% GMM (2 classes)<br>75.1% GMM (3 classes)<br>94.9% RF (2 classes)<br>91% RF (3 classes) | 20 |
| Barki et al. (2023)[54] | MAX30102 BNO055 | PPG, ACC | CNN | Stroop, Mental Arithmetic | 92.04%<br>96.02% adding white noise | 14 |
| Mech et al. (2022)[55] | Samsung Gear S3 | HR, ECG | Statistical | Math, Stepping | - | 52 |
| Seo et al. (2022)[56] | Zephyr | ECG, Resp | DNN | Stroop | 73.3% | 24 |
| Umer (2022)[57] | Equivital EQ02 | ECG, ST, Resp, GSR | Bagged Trees | Math, physical handling | 94.7% | 8 |
| Chalabianloo et al. (2022)[58] | 7 devices* | ECG, HR, GSR | ERT, RF | Stroop, cycling | 83.89% (E4, HR)<br>90.62% (E4, HR & GSR) | 32 |
| Li et al. (2022)[59] | midbeat | ECG | Statistical | Nurses workplace | - | 17 |
| Fauzi et al. (2021)[60] | Empatica E4 | GSR, ACC, ST, BVP | NB, SVM, RF, NN, K-NN, LR, DT | Hospital staff | 86.61% (RF)<br>87.10% (ensemble) | 15 |
| Dai et al. (2021)[61] | Fossil Gen4 | PPG, ACC | SVM, RF | Speech, Math, cold, daily life | 82.6% (laboratory)<br>79.8% (daily life) | 32 |
| A S et al. (2020)[62] | ADI-VSM | GSR | K-NN | Pre-surgery | 85.06% | 41 |
| Can YS et al. (2020)[63] | Empatica E4 | HRV,GSR, ACC | MLP, RF, LDA | Daily activities | 85.36 (3 classes)<br>98% (2 classes) | 16 |
| Kaczor et al. (2020)[64] | Empatica E4 | GSR, ACC, HR | DT, LR, NB, SVM, K-NN, DA | Physicians in the emergency department | 70% | 8 |
| Kyriakou et al. (2019)[65] | Empatica E4 | GSR, ST | New Algorithm | Audible, real-world urban | 84% | 11 |
| Lopez et al. (2019)[66] | Empatica E4 | GSR | Statistical | Stroop, audible | 79.17% | 12 |

Abbreviation: HR: Heart Rate; ECG: Electrocardiogram; Resp: Respiration; ST: Skin Temperature; GSR: Galvanic Skin Response; HRV: Heart Rate Variability; ACC: Acceleration; PPG: Photoplethysmogram; BVP: Blood Volume Pulse; DNN: Deep Neural Network; MLP: MultiLayer Perceptron; RF: Random Forest; LDA: Linear Discriminant Analysis; DA: Discrimination Analysis; ERT: Extremely Randomized Tree; SVM: Support Vector Machine; K-NN: K-Nearest Neighbor; NB: Naïve Bayes; LR: Logistic Regression; DT: Decision Tree; GMM: Gaussian Mixture Model; SOM: Self-Organizing Map; CNN: Convolutional Neural Network. *7 devices are BITalino (r)evolution board, Firstbeat Bodyguard2, Polar H10, Zephyr HxM, Empatica E4, Samsung Gear S2, and CoreSense.

# Chapter 5

# Materials and Methods

## 5.1 Introduction

This work's primary objective is to analyse data collected by Empatica E4 and to assess the validity of our model, using Machine learning techniques. The data have been collected in our laboratory after the candidates have been instructed about the protocol, the aim of this study, and signed privacy questionnaires. A total of 29 subjects were individuated and equipped with Empatica E4. The chosen environment for this study is Matlab 2022 for pre-processing, and Python for classification. This section is divided into subsections, namely Empatica E4, Data Acquisition Protocol, Data Pre-Processing, Features extraction, Machine Learning algorithms, and Deep Learning, each dedicated to a specific portion of the work carried out.

## 5.2 Materials

### 5.2.1 Empatica E4 bracelet

The Empatica E4 bracelet was the instrument used in the study. It is a wearable device made to continuously and instantly gather data. The temperature sensor, accelerometer, EDA sensor that measure the skin's galvanic impedance, and PPG sensor that allow the detection of the blood volume pulse make up the E4's four sensors, all of which are useful for the detection of physiological data Figure 5.1 shows the position of Empatica E4 sensors. For improved stability of the device on the wrist during testing, the subjects wore the E4 in such a way that it was snug enough not to slide along the arm.
The Bluetooth streaming acquisition was chosen to store the data. When connecting the E4 to a smartphone via the E4 Real-time app, the bracelet will automatically begin recording the physiological parameters, which the app can then monitor in real-time shown in Figure 5.2. Then, the data was immediately uploaded to E4 Connect, Empatica's cloud platform Figure 5.3, to download them and start with the pre-processing.


Figure 5.1: Empatica E4 with the position of its sensors [67].

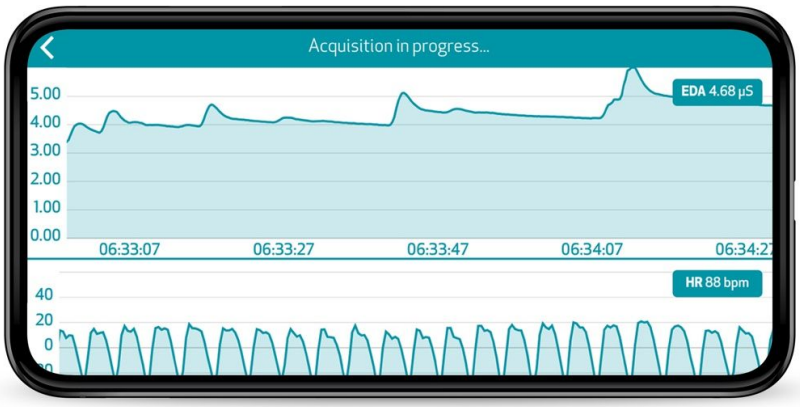



Figure 5.2: E4 mobile streaming interface [67].

### Empatica E4 Technical Specifications

- **PPG Sensor:** This sensor samples at a rate of 64 Hz and utilizes 4 light-emitting diodes (2 green and 2 red) along with 2 photodiodes to capture signals. Green light provides information on heartbeats, while red light helps reduce motion artifacts.
- **EDA Sensor:** The Electrodermal Activity (EDA) sensor measures skin electrical conductance changes at a rate of 4 Hz, within the range of 0.01 to 100 µS, with a resolution of 900 pS. It uses stainless steel (standard) or Silver (Ag) plated with a metallic core, electrodes placed on the wrist, and applies a small alternating current to the skin.
- **IR Thermometer:** Configured to sample at 4 Hz, this sensor measures skin temperature (SKT) using an optical thermopile sensor. It maintains accuracy


Figure 5.3: Empatica E4 Working Modes [67].

within the human skin temperature range (36–39 ◦C) with a calibration range of -40 to 115 ◦C.

- **3-Axis Accelerometer:** With a fixed sampling frequency of 32 Hz, this accelerometer provides high-sensitivity motion detection across three axes (X, Y, Z) and a default range of ±2g. Custom firmware allows for the selection of ranges ±4g or ±8g with a resolution of 8 bits.

### 5.2.2 Data Acquisition Protocol

In order to evaluate mental stress, a protocol must be defined and appropriate stressors should be identified. Several categories can be individuated, such as cognitive stressors, characterized by tasks that require a high level of attention, concentration, and memory, such as solving complex mathematical problems or memorizing long word lists. Social stressors can be perceived as threatening or judgmental, such as participating in a job interview or giving a public speech. Physical stressors are those situations that require intense physical effort, such as engaging in high-intensity exercise or being exposed to extreme temperatures. Finally, we can distinguish between emotional stressors, which can elicit intense and negative emotions, and psychological stressors, which require experiencing a sense of uncertainty or lack of control [68].
On the basis of the different stressors, we searched the literature in order to recreate a protocol that combined all these stressors to create more complex and realistic mental stress. We focused on creating mainly cognitive, social, and physiological stressors since they are the most likely to be triggered in a working environment and easily induce cognitive load in a laboratory situation.
Therefore, we came out with the protocol depicted in Figure 5.4, developed based on the one suggested in [69]. We decided to apply this protocol as a base to develop our own due to the high accuracy that the just quoted study reached.
Three minutes of rest were recorded after the bracelet was turned on in order to

establish a baseline. At the conclusion of each task, a 2-minute rest period was carried out. In the first task, participants had ten minutes to construct a Lego object using only the images printed on the box and no instructions. The second task is to assemble the same Lego creation within five minutes, but this time with the aid of the instructions. The third task requires the participant to assemble another Lego creation made of larger pieces in three minutes while following instructions and counting backward from 180 (the total amount of time available to complete the task) to zero. Each of the aforementioned tasks was developed to simulate manufacturing activities such as assembly and manual handling and to induce the mental stress that workers can face while doing a specific job. The fourth test is entirely mathematical and involves repeatedly subtracting backward the number 13 from 511. There is no time limit in this situation. This task is inspired by the Montreal Imaging Stress Task, created to investigate the effects of psycho-social stress in the human brain [70]. The fifth and final test requires the subject to give a one-minute oral presentation of themselves and their resume since it has been demonstrated that an oral presentation can cause stress and memory impairments [71].

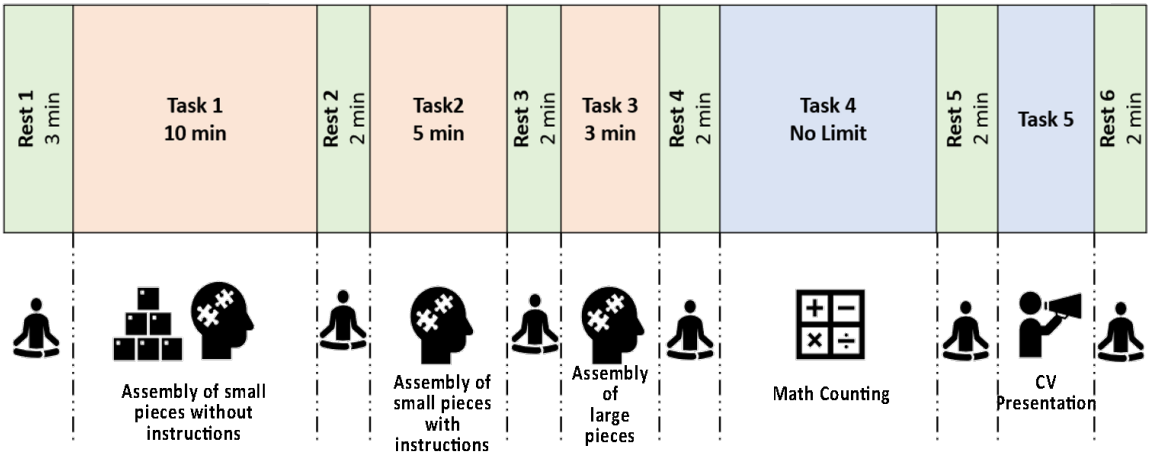



Figure 5.4: Data acquisition protocol carried out for each of the partecipant

## 5.3 Method

### 5.3.1 Machine Learning Approaches

#### Data pre-processing

After the data acquisition, the data have been filtered to extract features and apply Machine Learning algorithms.
The duration of the signal segment is known to affect HRV and pulse rate variability (PRV) features [72]. This implies that the HRV features are contingent upon the length of the segment under consideration. In this study, we employed two pre-processing

methodologies for segmenting the PPG and EDA signals into 1-minute segments. Several factors influenced the decision to use a 1-minute interval. Firstly, the data collection protocol for this study included a 1-minute task - the CV presentation task. Secondly, a one-minute duration is appropriate for use in wearable health monitoring devices. Thirdly, in order to maximize data segments:

1. 1-Minute Non-overlapping Segments: We initiated the segmentation of the PPG and EDA signals into intervals of 1-minute duration. The segmentation process included all 29 subjects who took part in this study. For each BVP and EDA signal, a total of 1068 data segments were extracted.

2. 1-Minute with 1-Second Sliding Window with 59-Second Overlapping Segments. This approach allows us to capture more segments. The segmentation process remains the same, but instead of non-overlapping 1-minute segments, we create segments that slide forward by 1 second, resulting in a 59-second overlap between consecutive segments. In this approach for each BVP and EDA signal, a combined total of 46,030 data segments were generated. The sliding window approach is a widely used method in the segmentation step in order to maximize the number of segments and make a consistent analysis of the signals under consideration [73][74][75]. Figure 5.5 illustrates the non-overlapping and overlapping windowing techniques.

Following that, the segments were labeled according to the tasks or rest periods. Regarding the PPG signals, different noises and artifacts can affect the signals during

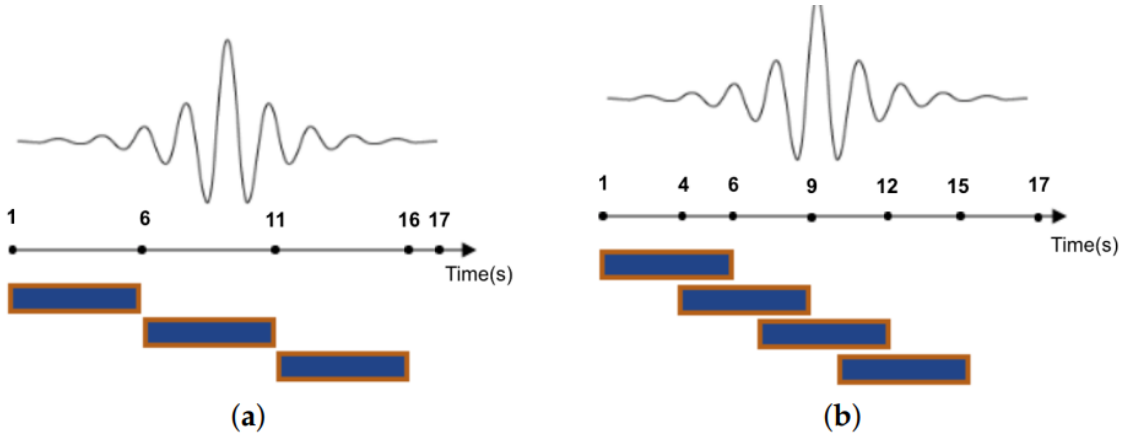
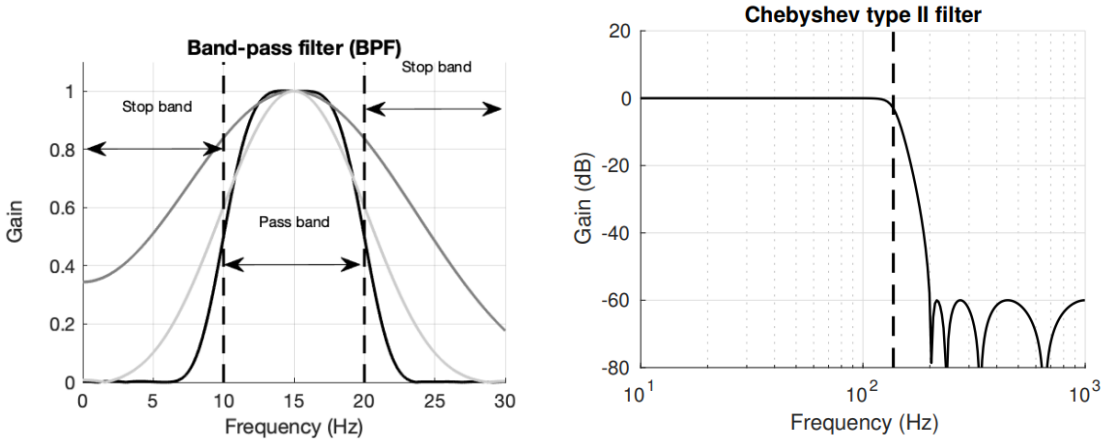



Figure 5.5: Sliding windows. (a) Non-overlapping; (b) Overlapping-2 s sharing [73].

PPG recording, lowering the stress detection system's accuracy. The most prevalent of them is the motion artifact, which has a significant impact on the PPG signal quality. For this reason, all the segments were filtered using a Chebyshev II order 4 filter with a stopband attenuation of 20dB and a passband of 0.5-5 Hz [76]. Figure 5.6




(a) (b)

Figure 5.6: (a) Frequency-domain behaviour of a band-pass filter(BPF). (b) Frequency response of a Chebyshev type II filter [77].

shows (a) frequency-domain behaviour of a band-pass filter(BPF). (b) Frequency response of a Chebyshev type II filter.

The crucial step is pinpointing the peak of the PPG signal and the distance between two consecutive peaks. Therefore, peak detection was performed using the *findpeaks* function, with a threshold set to a minimum peak distance of 0.4s and a minimum peak height of 0. Afterwards, peak-to-peak matrices were calculated by subtracting every two consecutive peaks. Following these computations, only intervals with a time duration of 500 to 1200 ms (corresponding to heart rates of 120 and 50 beats per minute) were taken into consideration, while all abnormal intervals (time duration less than 500 ms or greater than 1200 ms) were excluded. These limits were chosen based on the work of Zubair et al. [78] but we modified the lower limit to 500ms because it produces 120 Bpm instead of 600 ms which corresponds to 100 Bpm. This means that if we choose 600 ms, all HR more than 100 Bpm will be eliminated, removing HR values associated with stress tasks. However, excluding too many abnormal intervals would reduce the length of the PRV series. PPG segments with abnormal intervals that made up less than 15% of all intervals were therefore taken into account. The threshold of 15% was selected to ensure that the selected PPG segment still has a time length greater than the 50s after removing abnormal intervals [78]. As a result, the total number of PPG segments in a non-overlapping way was reduced to 843, obtaining 320 segments for the rest condition and 523 for all the tasks. While for overlapping way the total number of segments was reduced to 35,285, obtaining 10,439 segments for the rest condition and 24,846 for all the tasks.

For what concerns EDA pre-processing, upsampling from 4 to 64 Hz was performed to make both signals at the equal sampling frequency [79]. To remove any artefacts, smoothing using the Gaussian low pass filter, with a 40-point window and sigma of 400 ms, was carried out [38, 80]. Finally, all the clean segments went through the

feature extraction process. In Figure 5.7, there is a schematic representation of the PPG and EDA signal processing, while in Figure 5.8 5.9 are visible the signals before and after the cleaning.

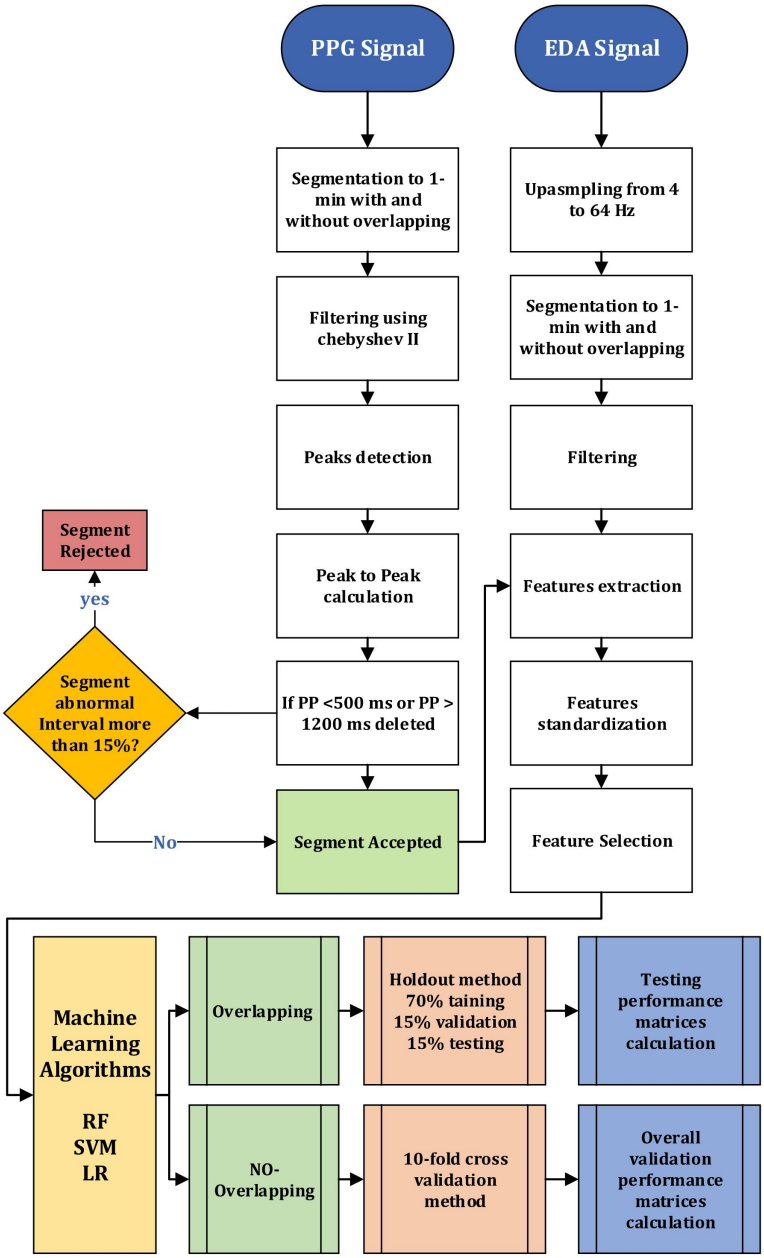



Figure 5.7: Flowchart for overall Machine Learning approaches, including PPG and EDA pre-processing.

### Features extraction and selection

Meaningful information was extracted from each data segment during the features extraction phase to characterize the various data portions in the time and frequency domains. Table 5.1 lists the 27 features that were chosen to quantify our data after being successfully applied in earlier studies for both PPG and EDA signals. For

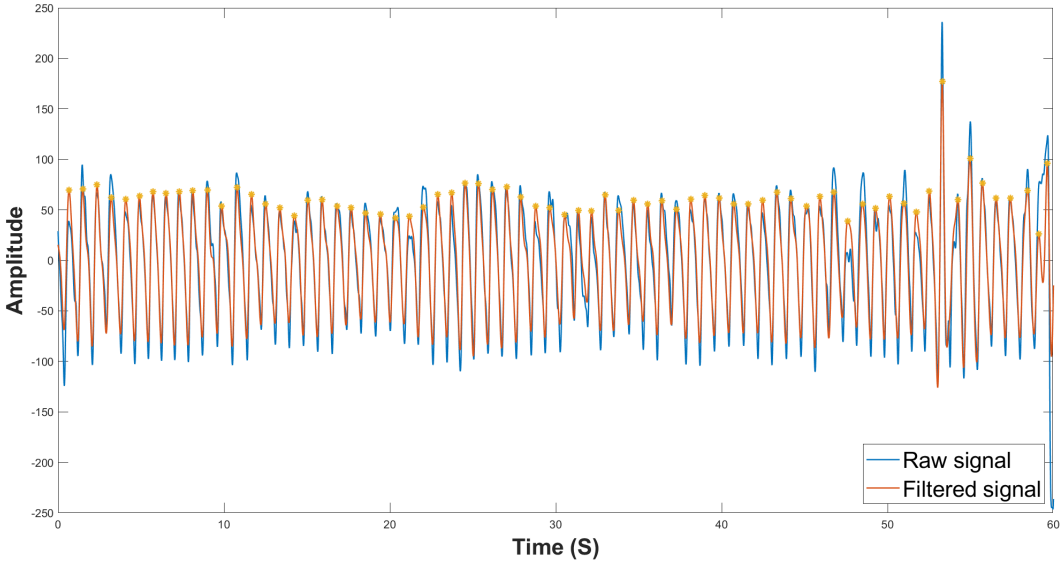



Figure 5.8: Raw and clean PPG signal.

the BVP, a total number of 16 features were extracted, in particular, the features could be divided into two categories: first one PRV based on calculated Peak-to-peak (PP) matrices and it is worth mentioning that only consistent features in ultra-short term matrices were included [72]. The second one is related to the signal itself such as mean, median, mode, minimum, maximum, standard deviation, mean and standard deviation of the first and the second derivative of the filtered signal [38, 79]. To extract this information, an algorithm was developed using several functions available on Matlab Statistics and Machine Learning Toolbox. Down below are reported the mathematical formula for the statistic features and the ones computed in the frequency domain. Equation 5.1 shows the formula for the mean computation, while Equations 5.2, 5.3 represents the median and standard deviation, respectively. In Equation 5.4 the absolute power in high frequency is reported, where $f(\lambda)$ is the power spectrum of the PP tachogram [81]. Finally, Equation 5.5 is based on the summation of successive PP intervals like a moving average. Its deviation represents the "long term HRV" [81].

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5.1}$$

$$\begin{cases} x_{(k+1)/2} & n \text{ odd} \\ \frac{x_{n/2}+x_{n/2+1}}{2} & n \text{ even} \end{cases} \tag{5.2}$$

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{5.3}$$

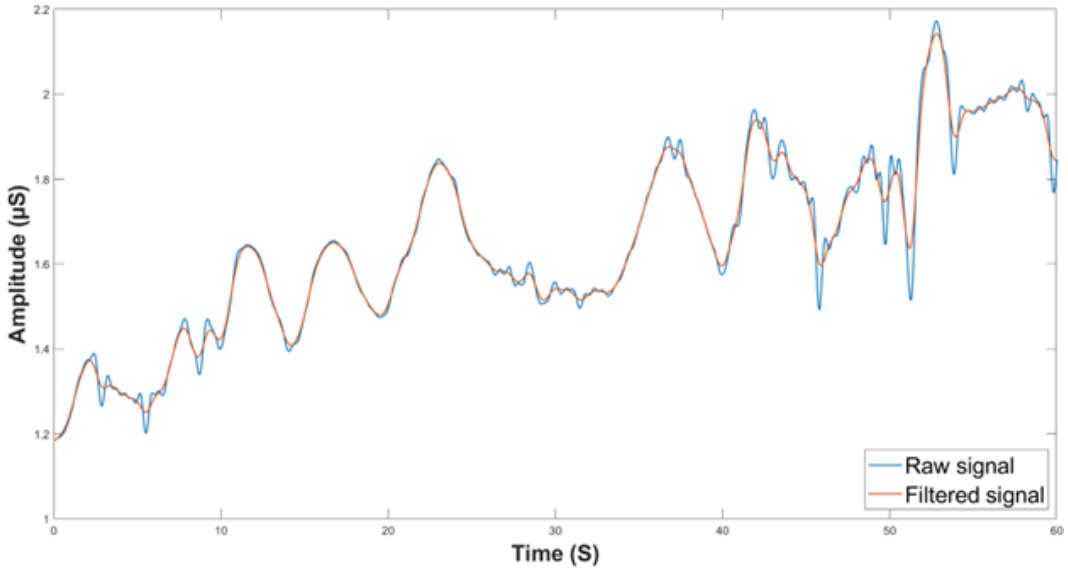



Figure 5.9: Raw and clean EDA signal.

$$HF = \int_{0.15\text{Hz}}^{0.40\text{Hz}} f(\lambda)\, d\lambda \tag{5.4}$$

$$SD2 = \sqrt{\frac{1}{2}} \cdot \text{std}(PP_{i+1} + PP_i) \tag{5.5}$$

For the maximum and minimum values of each signal, the functions *min[x]* and *max[x]*, with $x$ as a signal, from the previously mentioned toolbox were applied.
The BIO-SP tool was used to extract skin conductance response (SCR) features. SCRs are commonly found in electrodermal activity signals and can be identified using differentiation and convolution with a 20-point Bartlett window. This method is commonly used in EDA signal analysis to identify and characterize SCRs, which are important indicators of sympathetic nervous system activity. All of the features available in this tool were extracted, including the mean rise time, duration, amplitude, number of peaks, and mean of the SCR signal [80]. In the end, a feature standardization using the Z-score was performed since the parameter magnitudes were different.
In order to enhance the effectiveness of stress detection, the most pertinent and significant features should be chosen. The ranking of feature importance was performed using two methods: Univariate feature ranking for classification using chi-square tests (Chi-test), in Matlab, and the Pearson's correlation coefficient with the Waikato Environment for Knowledge Analysis (WEKA) [36]. The detailed explanations of both methods can be found in chapter 3.
In our case, using the Matlab function *fscchi2*, which examines whether each predictor variable is independent of a response variable by using individual chi-square tests. A small p-value of the test statistic indicates that the corresponding predictor variable is dependent on the response variable, and, therefore is an important feature. We

computed the predictor scores as –log($p$), with $p$ being the p-value. Therefore, a large score value indicates that the corresponding predictor is important. Then, we computed the mean value of the score and used it as a threshold.
Then, using WEKA, the function *CorrelationAttributeEval* was applied to evaluate the worth of an attribute by measuring the correlation between it and the class using Pearson correlation coefficient. Any attributes with rankings below a cutoff of 0.10 were eliminated [38].

### Classification

A class label related to the presence or absence of stress is returned from the ML classifiers using the subset of features produced, as well as the total set of features, as input. Based on the literature review reported in Chapter 4, the most frequently used and effective binary classifiers for identifying stress have been implemented. In particular, Random Forest and Logistic Regression, and SVM with cubic kernel in Python scikit-learn.
For what concerns the non-overlapping way since we have limited data size, these three approaches were tested using a 10-fold cross-validation configuration setting to test the Machine Learning algorithms for the model evaluation. In this configuration, the new features dataset was divided into 10 subsamples randomly, with 9 subsamples serving as training data and 1 subsample serving as validation data. The resulting accuracy percentage is the average over the 10 iterations using the available subsamples as validation data. Instead in the overlapping way since we have more data segments, the holdout configuration was applied. In particular 70% for testing, 15% for validation, and 15% for testing. In this configuration model evaluation was accessed using the testing part. Figure 5.7 illustrates the overall flow chart for machine learning approach.
The tuning of hyperparameters was done for all three classifiers in both ways using Optuna framework which is described in chapter 3. In order to get the best metrics, to optimize a Random Forest classifier following hyperparameter ranges have been applied:

- 'n estimators': we vary the number of trees in the forest within the range of 50 to 200, evaluating different ensemble sizes.
- 'max depth': the maximum depth of each tree is adjusted, with values ranging from 2 to 20, allowing the trees to capture different levels of complexity.
- 'min samples split': we investigate the minimum number of samples required to split an internal node, ranging from 2 to 10, controlling tree branching.
- 'min samples leaf': the minimum number of samples required to be at a leaf node is explored, with values ranging from 1 to 10, influencing leaf node granularity.

Table 5.1: All the features computed with their domain and abbreviation.

| Signal | Domain | Features | Abbreviation |
|---|---|---|---|
| PPG | Time | Mean PPI, standard deviation of PP interval, mean heart rate, standard deviation of heart rate. | Mean_PP, std_PP, M_HR, std_HR |
| | Frequency | Absolute power in high frequency [0.15-0.4 Hz]. | HF |
| | Non-Linear | Heart long-term variability. | SD2 |
| | Statistical | Mean of the filtered signal, median of the filtered signal, mode of the filtered signal, minimum of the filtered signal, maximum of the filtered signal, the standard deviation of the filtered signal, mean of the first derivative of the filtered signal, the standard deviation of the first derivative of the filtered signal, mean of the second derivative of the filtered signal, the standard deviation of the second derivative of the filtered signal. | Mean_BVP, Median_BVP, Mode_BVP, Min_BVP, Max_BVP, Std_BVP, M_d1, Std_d1,M_d2,Std_d2, |
| EDA | Statistical | Mean, median, mode, maximum, minimum, standard deviation. | Mean_EDA, Median_EDA, Mode_EDA, Max_EDA, Min_EDA, Std_EDA |
| SCR | Time | Mean Duration, Mean Amplitude, Mean Raise Time, Number of peaks, Mean. | M_D,M_Amp,M_RT,N_PEAKS, M_SCR |

In order to achieve the best results for the SVM model, the cubic kernel was used, and 'C': the regularization parameter was optimized within a logarithmic range from 1e-3 to 1e5. This range covers a broad spectrum of regularization strengths, allowing us to find the best trade-off between model complexity and fitting the data. For the Logistic regression model the 'C' hyperparameter was fine-tuned, representing the inverse of regularization strength. A logarithmic search space from 0.001 to 1000 and 'l2' regularization with the LBFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) solver were employed to optimize the model's performance and ensure its generalization capability.
The ability to categorize the presence or absence of stress, as a binary classification task, was assessed using the classification performance metrics of Accuracy.
Additionally, SHapley Additive exPlanations (SHAP) were employed in the top-performing model, Random Forest. This was done to establish a comparison with features selected through Pearson and chi-square methods and to gain insights into which specific features hold the most significant influence over the model's output.

### 5.3.2 Deep Learning Approaches

The proposed deep learning architecture is implemented in MATLAB to analyze stress presence using PPG signal. To train the models, various pre-trained Convolutional Neural Networks (CNNs) are employed, including GoogLeNet, and SqueezNet. The 1068 PPG segments, particularly (377 segments for rest, and 691 segments for stress) obtained previously from the non-overlapping way were filtered using a Chebyshev II order 4 filter with a stopband attenuation of 20dB and a passband of 0.5-5 Hz. Furthermore, the PPG signal is transformed into a time-frequency domain using CWT to obtain the wavelet coefficients (a scalogram). These scalograms, serving as intricate visual representations of the PPG signal, were depicted with logarithmic frequency scaling to facilitate meaningful interpretation. Then these scalograms were converted into RGB images[82][54]. The dimensions of these images were set to 224x224x3 and to 227x227x3 to ensure compatibility with the GoogLeNet and SqueezNet architectures respectively. Furthermore, the aforementioned networks parameters were modified to fit our aim of classification of the presence or absence of stress, followed by training using 80% training and 20% validation splitting configuration. Regarding the network's training, the parameters were set as follows:

- Optimization algorithm: Stochastic Gradient Descent with Momentum (SGDM), a common optimization algorithm used in deep learning.
- MiniBatchSize=10
- Max Epochs=15
- Initial Learn Rate=1e-4 for GoogLeNet and 3e-4 for SqueezNet.

- Validation Frequency=10:

Afterwards, key performance metrics, including accuracy, precision, recall, and F1-score, relevant to the specific classification class, were calculated and included in the results. Figure 5.10 illustrates the overall flow chart of the DNN approaches.

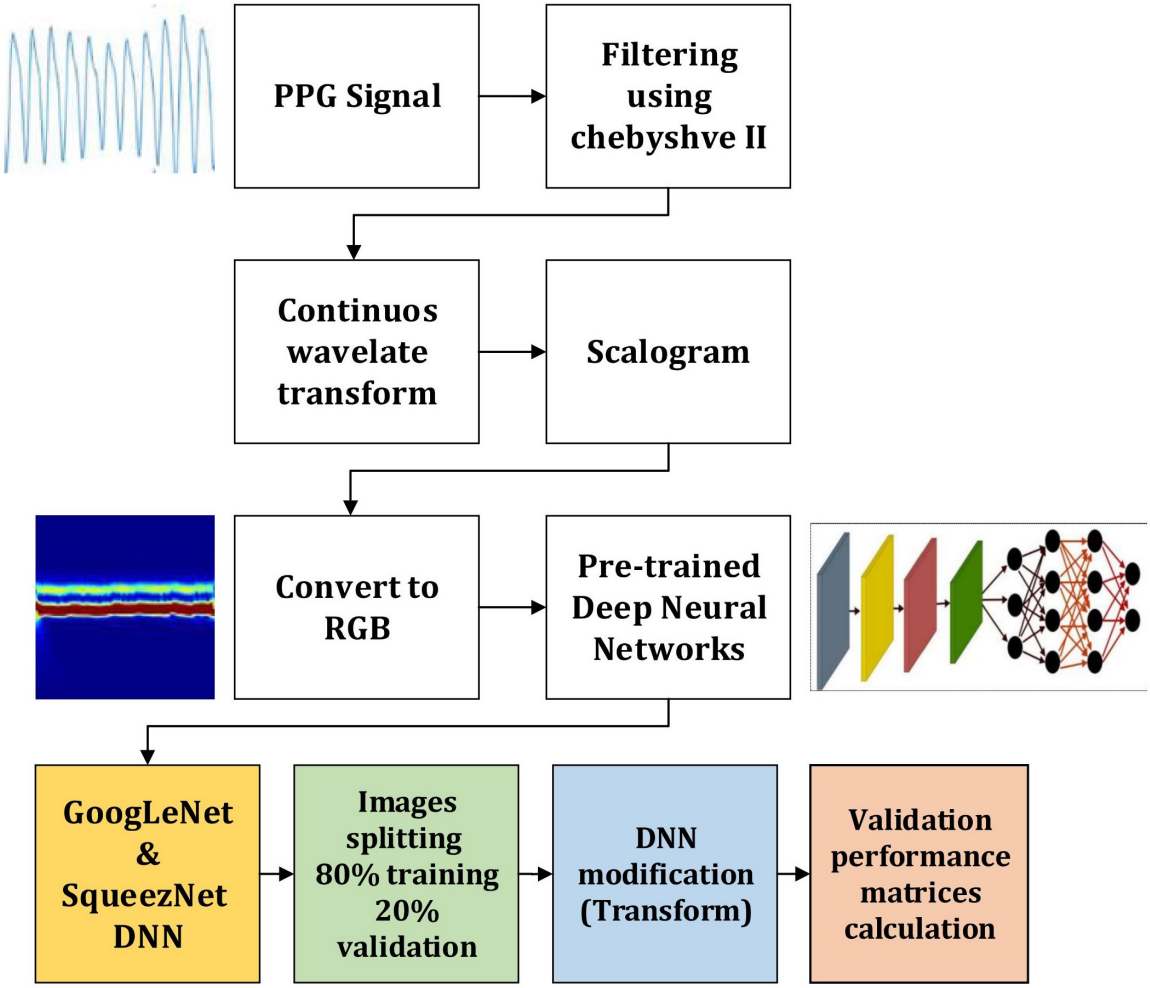



Figure 5.10: Flow chart of DNN approaches.

# Chapter 6

# Results

## 6.1 Feature selection

Empatica E4 data were pre-processed and analyzed to extract features from each recording. The features chosen for feeding the ML algorithms are then reported for both methods, in Figure 6.1 and 6.2. Applying the Chi-test method, only 10 features were chosen from the original 27 ones while using the Pearson correlation coefficient only 15 features fed the ML algorithms. Through both methods, it can be clearly seen that all HRV features exhibited values above the selected threshold in Pearson's correlation method. Additionally, in the Chi-square method, 4 out of the HRV features surpassed the threshold confirming the validity and stability of the information that they carry.

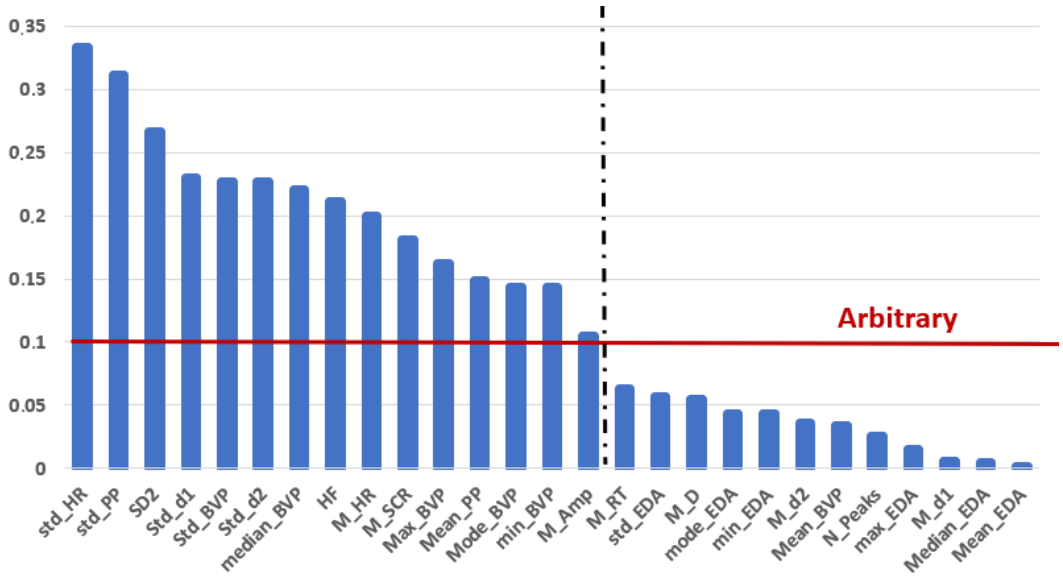



Figure 6.1: Ranks listed in order of importance for each feature extracted using Pearson's correlation coefficient.

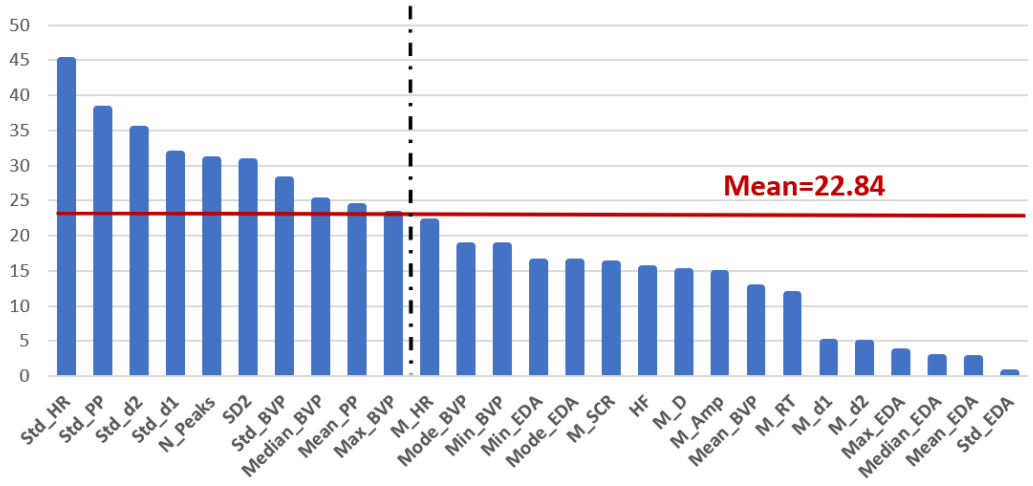



Figure 6.2: Ranks listed in order of importance for each feature extracted using Chi-test method.

## 6.2 Machine learning approaches

Using different machine learning algorithms, we assessed the performance of our system in classifying stress. We utilized three classifiers: Random Forest, Support Vector Machine (SVM), and Logistic Regression. The classification results are summarized in Tables 6.1 (without overlapping) and 6.2 (with overlapping). Notably, Random Forest consistently outperformed the other classifiers, with SVM and Logistic Regression also achieving reasonable results. Moreover, to validate the feature extraction process, we also fed the algorithms with all 27 features. Confusion matrices for all three classifiers before and after feature selection methods are plotted in Figure 6.3 in non-overlapping case, and Figure 6.4in overlapping case. Furthermore, in Figure 6.5, a bar plot is established to give an overview of all classifiers' Accuracy in all cases mentioned above.

Table 6.1: Performance metrics before and after applying the Chi-test and Pearsons's correlation coefficient methods for all the three Machine Learning techniques. **Case 1: without overlapping.**

| ML Algorithm | Chi-Test method | | | | | Pearson's Correlation Coefficient | | | | | All features | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | Label | Prec | Rec | F1 | Accuracy | Label | Prec | Rec | F1 | Accuracy | Label | Prec | Rec | F1 |
| Random Forest | 74.7% | 0 | 0.69 | 0.60 | 0.64 | 74.9% | 0 | 0.71 | 0.57 | 0.63 | 76.4% | 0 | 0.74 | 0.59 | 0.66 |
| | | 1 | 0.78 | 0.83 | 0.80 | | 1 | 0.77 | 0.86 | 0.81 | | 1 | 0.78 | 0.87 | 0.82 |
| SVM | 72.7% | 0 | 0.68 | 0.52 | 0.59 | 72.8% | 0 | 0.68 | 0.53 | 0.60 | 74.9% | 0 | 0.72 | 0.55 | 0.62 |
| | | 1 | 0.74 | 0.85 | 0.79 | | 1 | 0.75 | 0.85 | 0.80 | | 1 | 0.76 | 0.87 | 0.81 |
| Logistic Regression | 71.7% | 0 | 0.66 | 0.51 | 0.57 | 72.7% | 0 | 0.68 | 0.52 | 0.59 | 75.3% | 0 | 0.72 | 0.56 | 0.63 |
| | | 1 | 0.74 | 0.84 | 0.79 | | 1 | 0.75 | 0.85 | 0.80 | | 1 | 0.77 | 0.87 | 0.82 |

We observed that data overlapping significantly improved classification performance for all three classifiers, leading to higher accuracy, precision, recall, and F1 scores. Random Forest, in particular, demonstrated robustness in handling imbalanced data, consistently achieving better results. The choice of feature selection method (Pearson's correlation coefficient or Chi-square) did not significantly impact model

Table 6.2: Performance metrics before and after applying the Chi-test and Pearsons's correlation coefficient methods for all the three Machine Learning techniques. **Case 2**: **with overlapping.**

| ML Algorithm | Chi-Test method | | | | | Pearson's Correlation Coefficient | | | | | All features | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | Label | Prec | Rec | F1 | Accuracy | Label | Prec | Rec | F1 | Accuracy | Label | Prec | Rec | F1 |
| Random Forest | 98.4% | 0 | 0.99 | 0.95 | 0.97 | 99.1% | 0 | 0.99 | 0.97 | 0.98 | 99.5% | 0 | 0.99 | 0.98 | 0.98 |
| | | 1 | 0.98 | 1 | 0.99 | | 1 | 0.99 | 1 | 0.99 | | 1 | 0.99 | 1 | 0.99 |
| SVM | 80.2% | 0 | 0.82 | 0.41 | 0.55 | 83.3% | 0 | 0.86 | 0.50 | 0.63 | 91.4% | 0 | 0.93 | 0.76 | 0.84 |
| | | 1 | 0.80 | 0.96 | 0.87 | | 1 | 0.83 | 0.97 | 0.89 | | 1 | 0.76 | 0.87 | 0.81 |
| Logistic Regression | 75.2% | 0 | 0.63 | 0.36 | 0.46 | 75.7% | 0 | 0.66 | 0.40 | 0.50 | 79.4% | 0 | 0.73 | 0.47 | 0.57 |
| | | 1 | 0.78 | 0.91 | 0.84 | | 1 | 0.78 | 0.91 | 0.84 | | 1 | 0.81 | 0.93 | 0.87 |

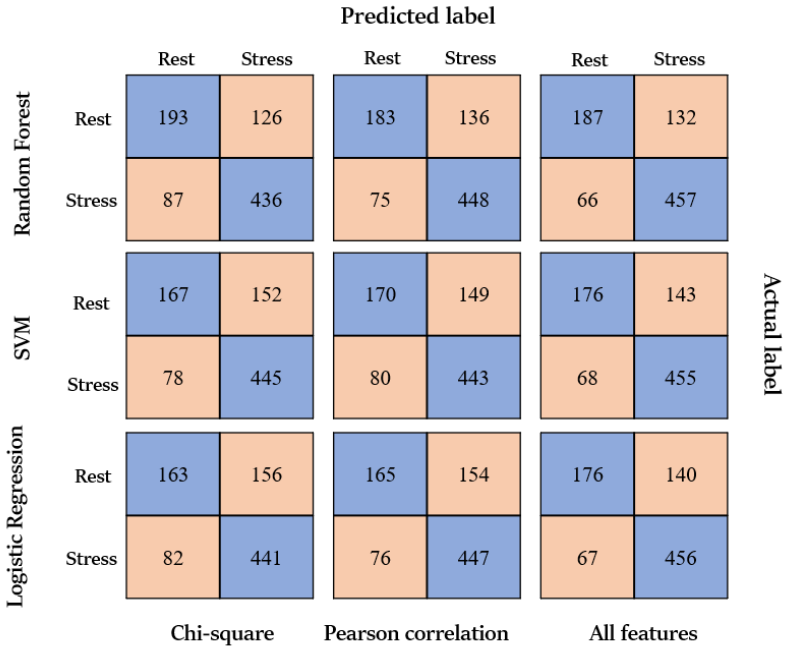



Figure 6.3: Validation confusion matrices for all the three Machine Learning techniques and before and after the features' selection. **Case 1: without overlapping.**

performance, demonstrating the robustness of our feature selection process.

### Random Forest SHeply exPlainability

To gain insights into feature importance and model decision-making, we employed SHAP values, as illustrated in Figures 6.6 and 6.7. These plots revealed that HRV-related features, such as Std HR and Std PP, played a crucial role in predicting stress, consistent with the results of our feature selection methods.

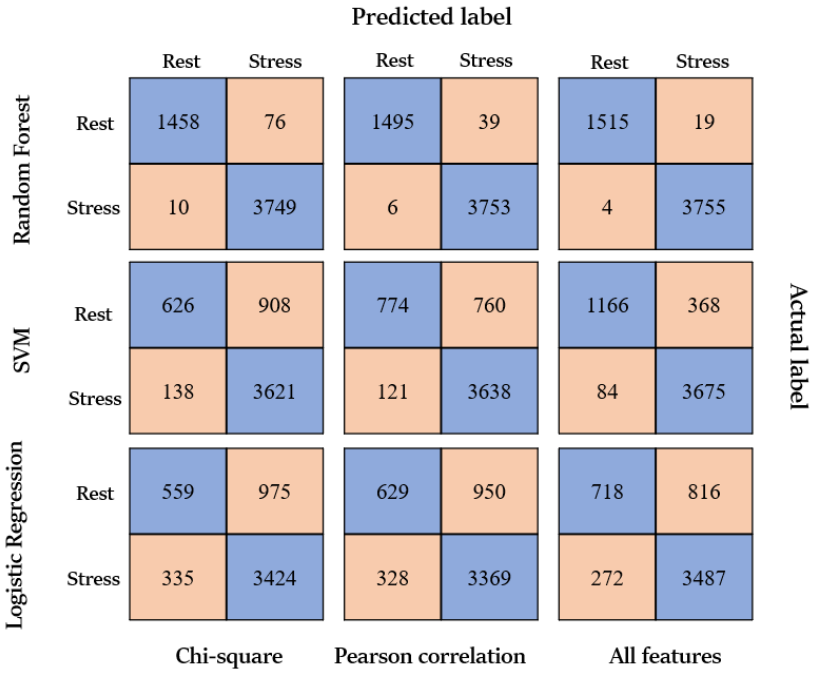



Figure 6.4: Testing confusion matrices for all the three Machine Learning techniques and before and after the features' selection. **Case 2: with overlapping.**

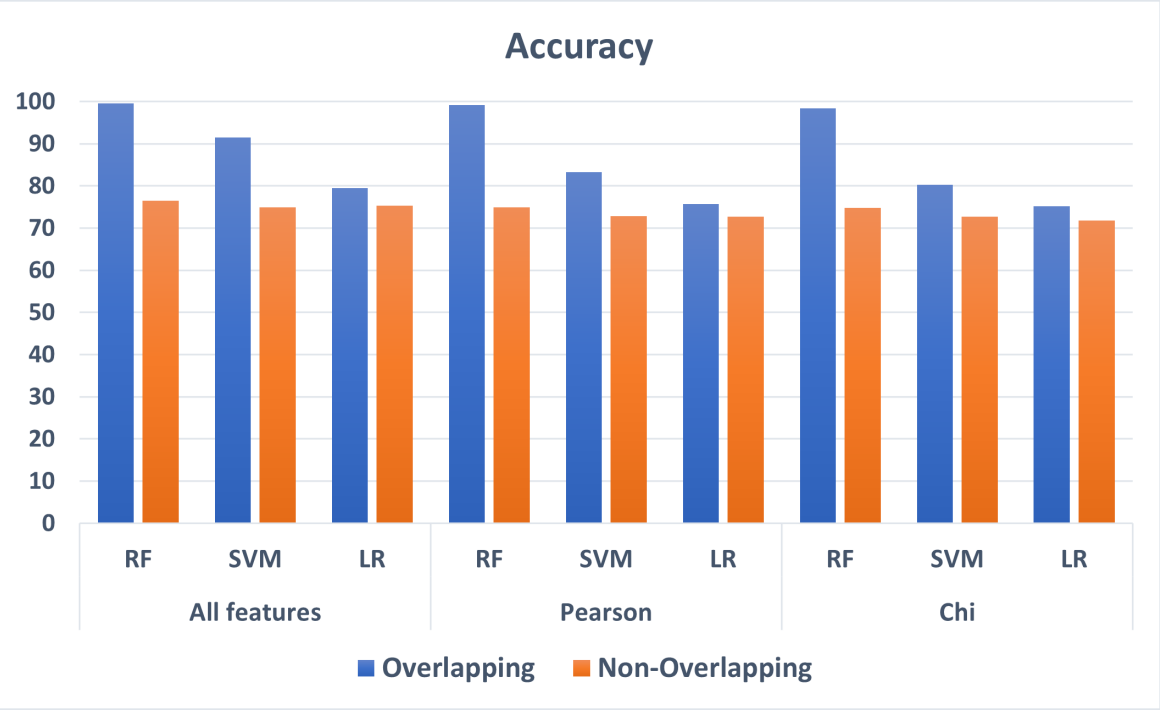



Figure 6.5: Bar plot of the accuracy before and after applying the Chi-test and Pearsons's correlation coefficient methods for all the three Machine Learning techniques in both overlapping and non-overlapping cases.

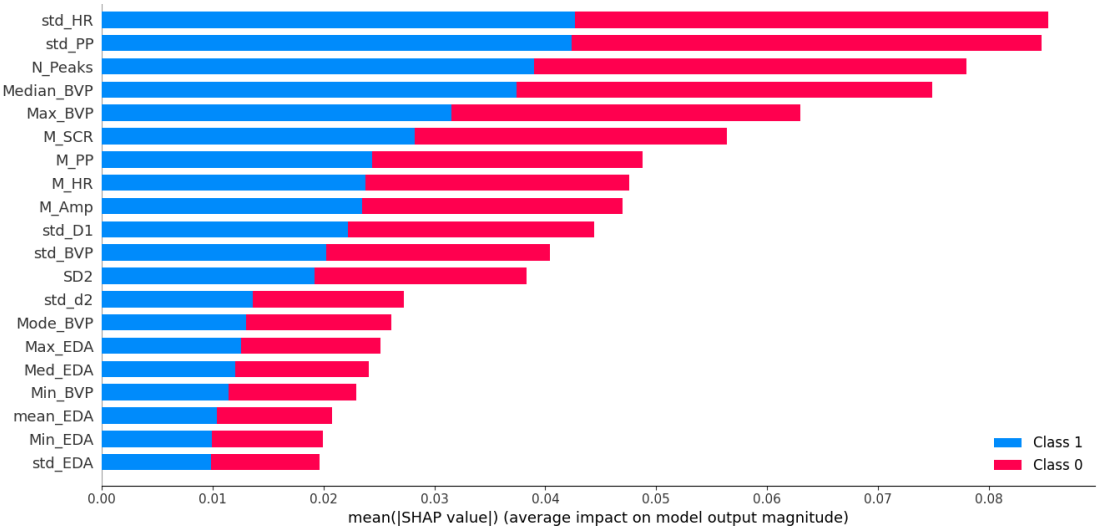



Figure 6.6: Feature influences with SHAP on both classes, with Random Forest model

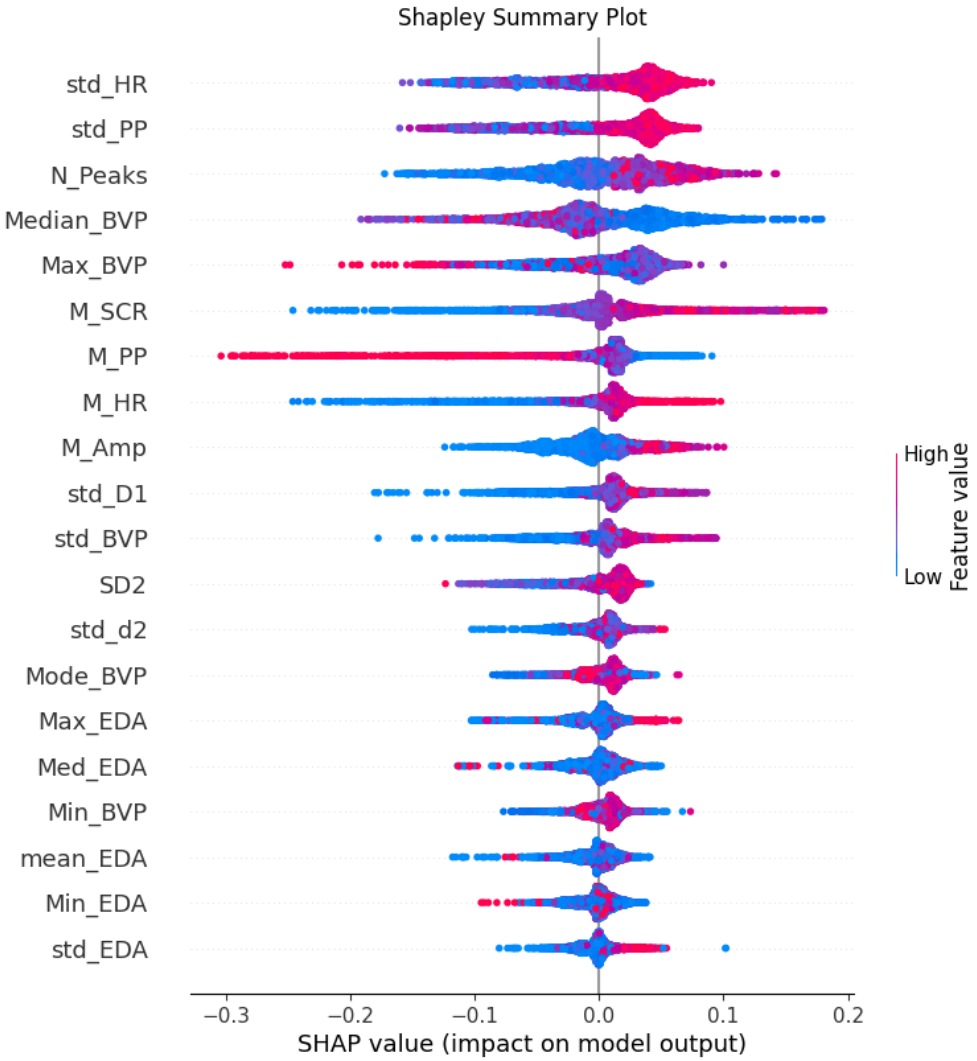



Figure 6.7: Feature influences with SHAP for the Stress class, with Random Forest classifier.

## 6.3 Deep Learning approaches

Lastly, we introduced a deep learning approach using convolutional neural networks (CNNs). Results of filtering and CWT are shown in Figures 6.9 and 6.10 for rest and stress, respectively. Results for GoogLeNet and SqueezNet models are presented in Table 6.3, confusion matrices in Figure 6.8, and the training graph for GoogLeNet and SqueezNet in Figures 6.11, and 6.12, respectively. These models achieved reasonable accuracy, precision, recall, and F1 scores, demonstrating their potential in stress detection.

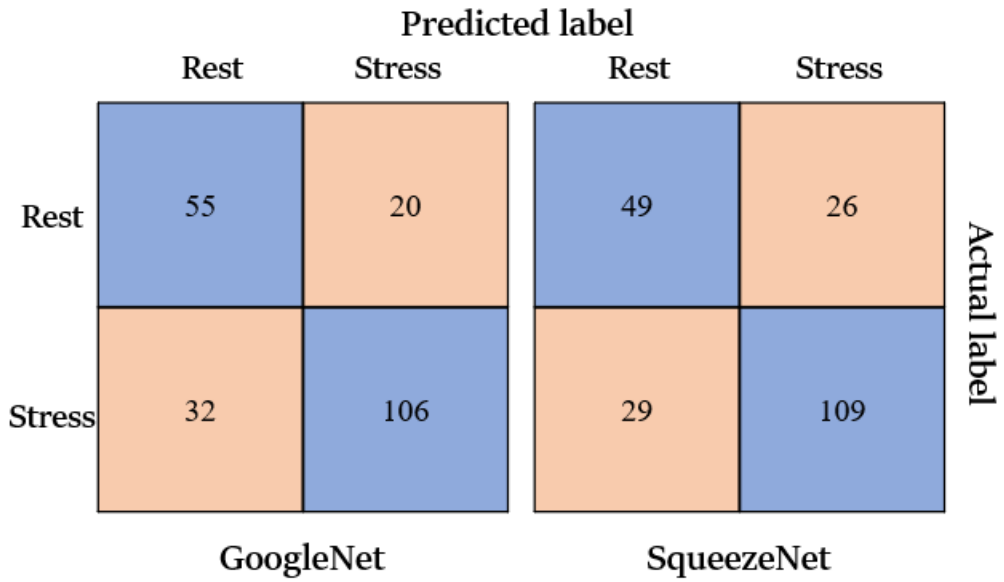



Figure 6.8: Validation confusion matrices for convolutional neural networks (CNN).

Table 6.3: Performance metrics for convolutional neural networks (CNN).

| CNN | Accuracy | Label | Prec | Rec | F1 |
|---|---|---|---|---|---|
| GoogLeNet | 75.6% | 0 | 0.63 | 0.73 | 0.68 |
| | | 1 | 0.84 | 0.77 | 0.80 |
| SqueezNet | 74.2% | 0 | 0.63 | 0.65 | 0.64 |
| | | 1 | 0.81 | 0.79 | 0.80 |

Overall, our study provides promising insights into stress detection using physiological signals from wearable devices. The system's performance is enhanced by data overlapping, and Random Forest emerges as a robust classifier. Feature selection methods and deep learning approaches further contribute to the accuracy of stress detection.

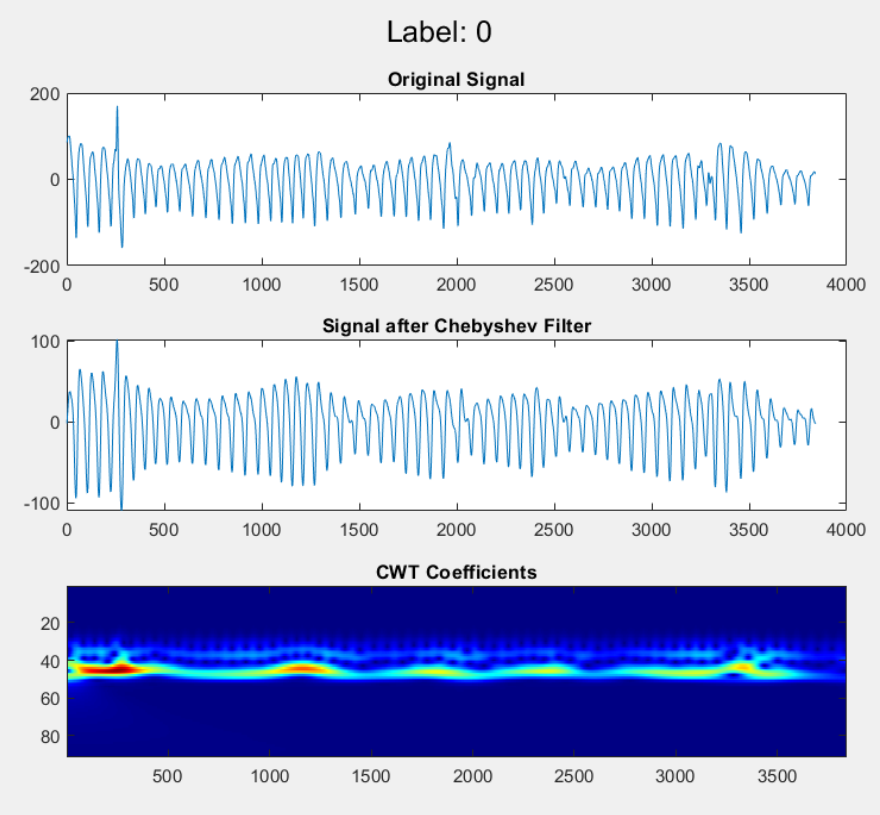



Figure 6.9: Original, filtered signal, and wavelet coefficients (scalogram) for Rest label segment.

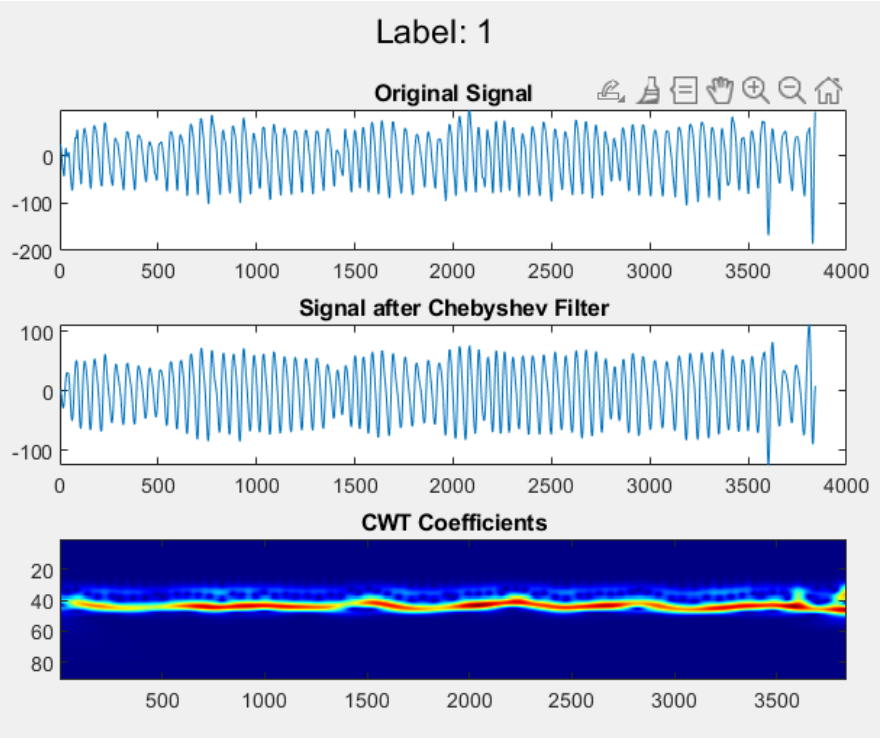



Figure 6.10: Original, filtered signal, and wavelet coefficients (scalogram) for Stress label segment.

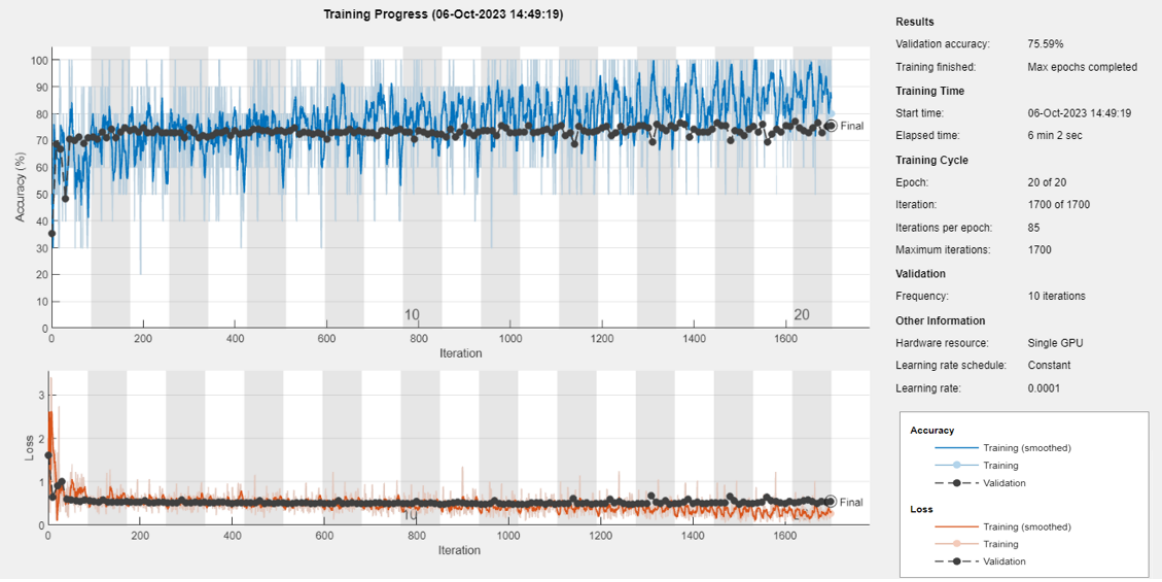



Figure 6.11: The training graph of pre-trained GoogLeNet CNN.

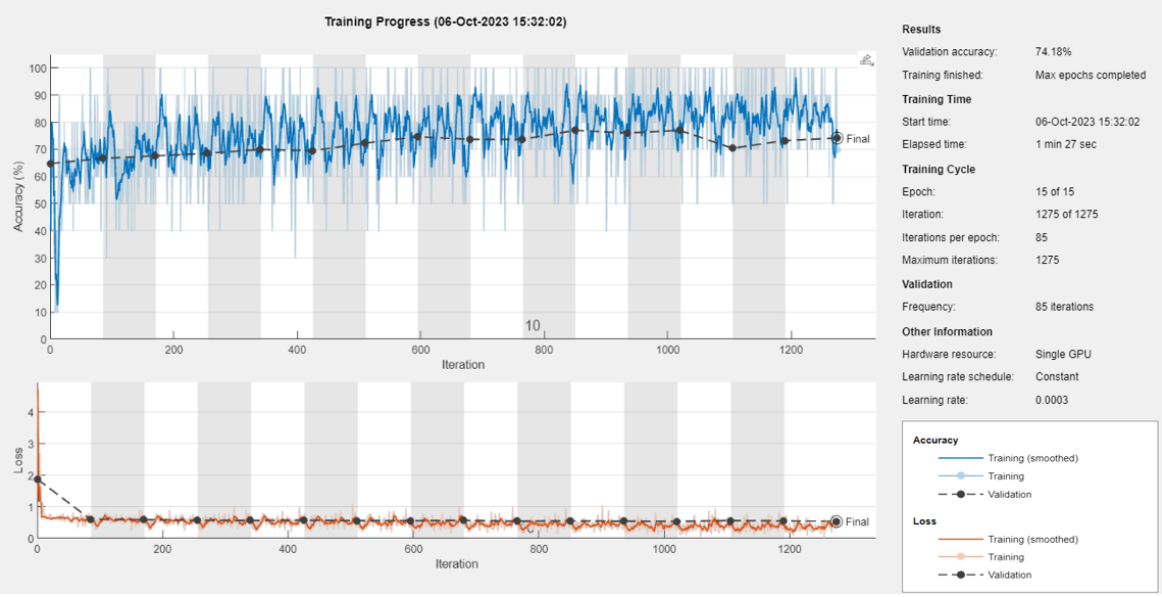



Figure 6.12: The training graph of pre-trained SqueezNet CNN.

# Chapter 7

# Discussion

A new system was proposed to analyze physiological signals measured with a wearable device on a test population before and after performing tasks designed to induce mental stress.

## 7.1 Feature selection

The Pearson coefficient was used for feature selection, and the results showed that most of the features were related to the PPG signal, while only two were related to the SCR signal. The top three features were the standard deviation of HR, PP, and long-term variability (SD2), respectively, with a ranking of over 0.25, indicating the importance of HRV analysis in detecting stress, consistent with previous research. For the BVP signal, the standard deviation of the first derivative, the standard deviation of the signal itself, and the standard deviation of the second derivative were the next three important features, respectively, with a ranking of over 0.1, suggesting that dispersion is more important than average values. As for EDA, only the mean of SCR and the mean amplitude of SCR, which are related to the phasic component of the EDA signal, had a rank higher than the threshold.
In the Chi-square approach, among the six HRV features, four were found to be above the predetermined threshold. Moreover, the standard deviation of PP and the standard deviation of HR were identified as the two most significant features. Regarding the BVP signal, similar to the Pearson correlation approach, the standard deviation of the second derivative, first derivative, and the signal itself were identified as important features. However, their mean values were not found to be significant. In the case of EDA signal analysis, it was found that only the number of SCR peaks was important with respect to the threshold. This observation confirms that the number of peaks (N_Peaks) is the primary indicator of Sympathetic Nervous System (SNS) activity, and, thus, related to a stress condition [83].
The findings of the current study are consistent with the existing literature in the field. Specifically, the results of the feature selection are aligned with the other studies that have emphasized the importance of the selected features. It is noteworthy that despite the differences in the devices employed in the previous studies, the HRV-

based features have emerged as the most robust indicator of stress, along with the SCR information [38, 55].
The results obtained from the current study indicate that both feature evaluation methods employed, particularly the Chi-test method, possess considerable strength in selecting stress-related characteristics. In fact, the outcomes achieved with the Chi-test method align with the ones obtained by [84]. Even if they applied different classifiers, the results are consistent with ours, demonstrating that the Chi-Test method is feasible for mental stress detection.

## 7.2 Machine Learning Approaches

Focusing on the classification aspect, in general, our analysis indicates that the classifiers' accuracy consistently exceeds 70%. This suggests that the pre-processing and original feature selection were appropriate for the database under consideration. Overall, the Random Forest algorithm consistently exhibited superior performance compared to other classification methods. This could potentially be attributed to the fact that Random Forest classifiers rely on randomness, which promotes more generalized modelling. This observation is corroborated by the Precision, Recall, and F1-measure metrics, which demonstrate the algorithm's effectiveness when contrasted with SVM and LR matrices. Observing the performance of the classifiers in detail, firstly without overlapping, Random Forest achieves an accuracy of 74.7%, 74.9%, and 76.4%, with Chi-Test, and Pearson's Correlation Coefficient, and all features methods, respectively. The F1-scores for both classes are decent but not extremely high. Secondly, with overlapping, the performance significantly improves, with an accuracy of 98.4%, and 99.1%, and 99.5% using Chi-Test, Pearson's Correlation Coefficient, and all features methods, respectively. The F1-scores are notably higher, indicating a significant enhancement in classification performance.
Analyzing the SVM classifier's performance, we first consider the results without overlapping. The accuracy achieved is 72.7% with the Chi-Test method, 72.8% with Pearson's Correlation Coefficient method, and 74.9% with all features set. The F1-scores for both classes show moderate performance. However, when data overlapping is applied, the classifier's performance improves substantially. It attains an accuracy of 80.2% with the Chi-Test method, 83.3% with Pearson's Correlation Coefficient method, and 91.4% with all extracted features. Moreover, the F1-scores for both classes experience a significant boost, indicating a substantial enhancement in classification performance.
Examining the performance of the Logistic Regression classifier in detail, we begin with the results obtained without overlapping. In this scenario, the classifier attains an accuracy of 71.7% with the Chi-Test method, 72.7% with Pearson's Correlation Coefficient method, and 75.3% with all features. While the F1-scores are reasonable, they do not reach exceptionally high levels, indicating a moderate classification

performance. Upon introducing overlapping, the classifier's performance remains relatively stable. The accuracy ranges from 75.2% to 79.4%. However, there are some notable improvements in the F1-scores, especially for class 1. This suggests that data maximizing has a more pronounced impact on the classifier's ability to differentiate class 1 (stress) instances.

The performance of all classifiers significantly improves when data overlapping is applied. This is particularly evident in terms of accuracy and F1-scores, indicating a substantial enhancement in classification performance. This suggests that data overlapping has a substantial positive impact on the ability of Machine Learning models to classify stress accurately. Due to the fact that data overlapping essentially increases the amount of training data available for classification.

Random Forest and SVM, in particular, benefit greatly from data overlapping, achieving much higher accuracy, precision, recall, and F1 scores. Logistic Regression also benefits from it, though the improvements are relatively smaller compared to the other two classifiers.

Observing the performance of the classifiers in detail, it becomes apparent that the classification of the presence of stress (label 1) outperforms the classification of its absence (label 0). This discrepancy may be primarily attributed to the unequal number of segments, as the number of segments associated with stress is greater than those associated with the rest phase. This is due to the fact that, in the overall protocol, the duration of the tasks is greater than the total rest period. Notably, class imbalance problem arises especially in SVM and Logistic Regression classifiers both with or without overlapping, which affects the overall performance. This is attributed to the fact that when imbalanced data are used to predict outcomes (by machine learning and data mining), the learning of the algorithm is affected. It is assumed that the data are drawn from the same distribution as the training data, presenting imbalanced data to the classifier and producing biased results [85]. Conversely, the Random Forest model outperforms in all situations, which is consistent with literature [86][87][88] due to its ability to handle imbalanced data in binary classification.

Despite the fact that the segments of rest and stress conditions were unbalanced, our results were still able to distinguish between these two conditions with a reasonable degree of accuracy particularly using Random Forest classifier. However, there is a possibility for improvement for example by balancing the data using different algorithms as the study carried out by [89] suggests. In fact, they obtained higher performance and accuracy after manipulating the data with ADASYN.

The choice of feature selection method (Chi-Test or Pearson's Correlation Coefficient) may not have a significant impact on model performance. Importantly, it does not result in a noticeable decrease in classification accuracy. However, it significantly influences the interpretability of selected features and contributes to computational efficiency by reducing dimensionality. Furthermore, this can also help prevent the model from overfitting [90].

Another factor to consider is the relatively small sample size used in our study.

Increasing the size and diversity of the participants would help to enhance the generalisability of our findings and improve the accuracy of our models. A small dataset may not be representative of the broader population and may be prone to inaccuracies and erroneous conclusions, as it could be influenced by outliers or anomalous data. Furthermore, the decision to conduct our study in a laboratory setting may have limited our ability to simulate real-world working conditions. In the future, additional stimuli could be introduced to overcome this limitation and more accurately replicate real-world scenarios.
Despite the drawbacks described above, our approach reached high performance in the detection the stress situations. This means that our choices for data manipulation and feature selection are sufficiently strong to deal with an unbalanced dataset. This means that in a real situation where motion artefacts have higher intensity and unpredictable stressful situations can arise, stress can be detected and monitored in order to avoid any psycho-physical complications.

## 7.3 Model explainability

Using stacked bar plots for global explainability, Figure 6.6 illustrates the mean SHAP values for all features. This represents the average influence of each feature on the output of the top-performing model (Random Forest), with overlapping approach from all features. The plot reveals that HRV time-domain features, specifically the standard deviation of Heart rate (Std HR) and standard deviation of peak to peak (Std PP), exhibit the most significant impact on the overall model output. Importantly, these features maintain the same ranking order in both feature selection methods applied (Pearson correlation coefficient and Chi-Test). The third most influential feature is the number of SCR peaks, a selection made based on the Chi-Test. Interestingly, noticing the remaining features, five out of the six HRV features also rank prominently in terms of impact. This consistency between the HRV feature importance and the results of the feature selection methods underscores the critical role of HRV features in stress detection models. Noticeably, the standard deviation of the BVP signal and its first and second derivatives show a good impact on the model output as observed in both applied feature selection methods.
To gain a more detailed understanding of how individual features impact each class (Local explainability), it's crucial to zoom in and examine the effects closely. Figure 6.7 presents horizontal scatter plots for each feature, utilizing distinct color gradients. This class-wise summary plot combines feature importances and feature effects. Each data point on the scatter plot represents a Shapley value for a feature and an observation. Features are positioned along the y-axis, while the Shapley values for their instances are displayed along the x-axis. To enhance visualization, overlapping points are slightly jittered. The color intensity and gradient in each instance indicate the feature values, ranging from low (blue) to high (pink), as illustrated in the color

bar.
Previously, we explored the behavior of Std HR and Std PP in Figure 6.6, where they emerged as the most impactful features in predicting the stress class. Now, we can delve into their behavior in a more detailed, class-specific manner. Upon closer examination of Figure 6.7, it becomes evident that higher values (leaning towards pink) of Std HR and Std PP tend to lead the model to classify the instance as "stress." Conversely, lower values (leaning towards blue) make it less likely for the model to predict "stress." In simpler terms, a higher presence of pink dots (indicative of high Std HR and Std PP values) corresponds to a stronger and positive effect (shift towards the right on the plot) on predicting the "stress" class. Similarly, we can observe that Mean Peak to Peak (M PP) is inversely correlated with the prediction of the "stress" class. This obvious correlation: as heart rate increases, there is typically a decrease in the peak-to-peak value (RR interval).
In summary, these plots allow us to gain an understanding of what our Machine-Learning model has learned from the features. Although there are variations in the feature importance rankings between the overall model output and the feature selection methods, there is a strong validation of the significance of the selected features.

## 7.4 Deep Learning Approaches

The objective of the proposed Deep Learning models was to employ deep transfer learning techniques for stress classification based on PPG signals. The overall performance assessment of the two suggested models indicates that the GoogLeNet model exhibits a slightly better performance than the SqueezNet model, achieving accuracies of 75.6% and 74.2%, respectively. Upon closer examination of the results, it is noteworthy that the models perform better at classifying the presence of stress (Label 1) compared to identifying its absence. This disparity could be attributed to the unequal distribution of segments, as there are more segments associated with stress than with the rest phase. Moreover, the choice of the Stochastic Gradient Descent with Momentum (SGDM) optimizer could potentially impact the performance of neural networks[54]. One limitation that affects the models is the relatively small sample size, which may restrict the generalizability of our results to a broader population. Furthermore, It is crucial to validate these results on different datasets and real-life applications to assess their generalizability. Despite the aforementioned limitations, achieving a 75.6% accuracy using only PPG signals demonstrates that using continuous Fourier transform and Convolutional Neural Networks (CNNs) holds promise as an effective approach for mental stress detection.

# Chapter 8

# Conclusions

## 8.1 Conclusion

The object of this study was to assess the stress presence by the measure of different physiological signals using a wearable sensor, Empatica E4. To address the stress condition a new acquisition protocol based on the literature reviews was used for 29 subjects. Moreover, the Machine Learning, and Deep Learning approaches were developed accordingly with the most performing algorithms in this field as well as the features chosen to characterize the model. Despite the limited baseline for the rest condition, which affected the balancing of the database. Among the classifiers we evaluated, Random Forest and SVM consistently yielded the best results, particularly when we applied data overlapping. With accuracies ranging from 80.2% to 99.5% and improved precision, recall, and F1-scores, these classifiers demonstrated their robustness in distinguishing stress from non-stress conditions. This was particularly noteworthy given the inherent challenges posed by an imbalanced dataset. Moreover, Deep Learning models exhibited potential in stress classification using PPG signals. To validate our model in the future, we suggest increasing the population size to include a diverse age range and implementing a new protocol that ensures a consistent baseline to avoid any misclassification issues. Additionally, incorporating multi-level stress tasks with different stressors, including real-life scenarios, could improve the model's robustness. It would be valuable to compare our results with feedback from participants, obtained through the use of questionnaires. Moreover, we recommend considering various wearable devices available on the market to assess the impact of their characteristics on the results.

I would like to acknowledge that a portion of the research presented in this thesis has been published in the following scientific article:
Campanella, Sara, Ayham Altaleb, Alberto Belli, Paola Pierleoni, and Lorenzo Palma. 2023. "A Method for Stress Detection Using Empatica E4 Bracelet and Machine-Learning Techniques" *Sensors* 23, no. 7: 3565. DOI: [https://doi.org/10.3390/s23073565]

# Bibliography


[1] Arthur C. Guyton and John E. Hall. *Textbook of Medical Physiology*. Elsevier Inc., Philadelphia, PA, 11th edition, 2006. International Edition ISBN: 0-8089-2317-X.

[2] Richard E Klabunde. *Cardiovascular Physiology Concepts*. Lippincott Williams & Wilkins, 2 edition, 2012.

[3] Shu Q. Liu. *Cardiovascular Engineering: A Protective Approach*. McGraw Hill, 1st edition, 2020.

[4] A. J. Weinhaus and K. P. Roberts. *Anatomy of the Human Heart*. Humana Press, New Jersey, 2005.

[5] Jaakko Malmivuo and Robert Plonsey. *Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields*. Oxford University Press, New York, 2002.

[6] Ioana-Raluca Adochiei, Felix Adochiei, Costin Cepisca, George Seriţan, Bogdan Enache, Florin Argatu, and Radu Ciucu. Complex embedded system for stress quantification. In *2019 11th International Symposium on Advanced Topics in Electrical Engineering (ATEE)*, pages 1–4. IEEE, 2019.

[7] Unai Zalabarria, Eloy Irigoyen, Raquel Martinez, Mikel Larrea, and Asier Salazar-Ramirez. A low-cost, portable solution for stress and relaxation estimation based on a real-time fuzzy algorithm. *IEEE Access*, 8:74118–74128, 2020.

[8] Dushyant Kumar Sharma. Physiology of stress and its management. *J Med Stud Res*, 1(001):1–5, 2018.

[9] Ayten Ozge Akmandor and Niraj K Jha. Keep the stress away with soda: Stress detection and alleviation system. *IEEE Transactions on Multi-Scale Computing Systems*, 3(4):269–282, 2017.

[10] Hymie Anisman and Zul Merali. Understanding stress: Characteristics and caveats. *Alcohol Research & Health*, 23(4):241, 1999.

[11] Sami Elzeiny and Marwa Qaraqe. Blueprint to workplace stress detection approaches. In *2018 International Conference on Computer and Applications (ICCA)*, pages 407–412, 2018.

[12] Muhammad Shahid Zafar, Museera Nauman, Hina Nauman, Sheema Nauman, Asifa Kabir, Zunaira Shahid, Anam Fatima, and Maria Batool. Impact of stress on human body: A review. *European Journal of Medical and Health Sciences*, 3(3):1–7, May 2021.

[13] Fátima González-Palau and Leonardo Adrián Medrano. A mini-review of work stress and mindfulness: A neuropsychological point of view. *Frontiers in Psychology*, 13, 2022.

[14] Fátima González-Palau and Leonardo Adrián Medrano. A mini-review of work stress and mindfulness: A neuropsychological point of view. *Frontiers in Psychology*, 13, 2022.

[15] Onur Parlak. Portable and wearable real-time stress monitoring: A critical review. *Sensors and Actuators Reports*, 3:100036, 2021.

[16] Martin Gjoreski and Mitja Luštrek. Matjaž gams, and hristijan gjoreski. 2017. monitoring stress with a wrist device using context. *Journal of Biomedical Informatics*, 73(10.1016).

[17] Giorgia Acerbi, Erika Rovini, Stefano Betti, Antonio Tirri, Judit Flóra Rónai, Antonella Sirianni, Jacopo Agrimi, Lorenzo Eusebi, and Filippo Cavallo. A wearable system for stress detection through physiological data analysis. In *Ambient Assisted Living: Italian Forum 2016 7*, pages 31–50. Springer, 2017.

[18] B Thanasekhar, N Gomathy, A Kiruthika, and S Swarnalaxmi. Machine learning based academic stress management system. In *2019 11th International Conference on Advanced Computing (ICoAC)*, pages 147–151, 2019.

[19] Davide Carneiro, Paulo Novais, Juan Carlos Augusto, and Nicola Payne. New methods for stress assessment and monitoring at the workplace. *IEEE Transactions on Affective Computing*, 10(2):237–254, 2017.

[20] Talha Iqbal, Adnan Elahi, Pau Redon, Patricia Vazquez, William Wijns, and Atif Shahzad. A review of biophysiological and biochemical indicators of stress for connected and preventive healthcare. *Diagnostics*, 11(3), 2021.

[21] Mengru Xue, Rong-Hao Liang, Bin Yu, Mathias Funk, Jun Hu, and Loe Feijs. Affectivewall: designing collective stress-related physiological data visualization for reflection. *IEEE Access*, 7:131289–131303, 2019.

[22] Edward Sazonov. *Wearable Sensors: Fundamentals, Implementation and Applications*. Elsevier Science & Technology, 2 edition, November 2020.

[23] Nannan Long, Yongxiang Lei, Lianhua Peng, Ping Xu, and Ping Mao. A scoping review on monitoring mental health using smart wearable devices. *Math. Biosci. Eng*, 19:7899–7919, 2022.

[24] Naghmeh Niknejad, Waidah Binti Ismail, Abbas Mardani, Huchang Liao, and Imran Ghani. A comprehensive overview of smart wearables: The state of the art literature, recent advances, and future challenges. *Engineering Applications of Artificial Intelligence*, 90:103529, 2020.

[25] Shruti Gedam and Sanchita Paul. A review on mental stress detection using wearable sensors and machine learning techniques. *IEEE Access*, 9:84045–84066, 2021.

[26] Jerry Chen, Maysam Abbod, and Jiann-Shing Shieh. Pain and stress detection using wearable sensors and devices—a review. *Sensors*, 21(4), 2021.

[27] Junyung Park, Hyeon Seok Seok, Sang-Su Kim, and Hangsik Shin. Photoplethysmogram analysis and applications: An integrative review. *Frontiers in Physiology*, 12, 2022.

[28] Ch Kiran kumar, M. Manaswini, K.N. Maruthy, A.V. Siva Kumar, and K. Mahesh kumar. Association of heart rate variability measured by rr interval from ecg and pulse to pulse interval from photoplethysmography. *Clinical Epidemiology and Global Health*, 10:100698, 2021.

[29] Sylvain Laborde, Emma Mosley, and Julian F. Thayer. Heart rate variability and cardiac vagal tone in psychophysiological research – recommendations for experiment planning, data analysis, and data reporting. *Frontiers in Psychology*, 8, 2017.

[30] Hugo F. Posada-Quintero, John P. Florian, Alvaro D. Orjuela-Cañón, and Ki H. Chon. Electrodermal activity is sensitive to cognitive stress under water. *Frontiers in Physiology*, 8, 2018.

[31] Nada Pop-Jordanova and Jordan Pop-Jordanov. Electrodermal activity and stress assessment. *PRILOZI*, 41(2):5–15, 2020.

[32] Ravi MD Bhoja, Oren T. MD Guttman, MPH Fox, Amanda A. MD, Emily MD Melikman, Matthew BS Kosemund, and MD Gingrich, Kevin J. MEngr. Psychophysiological stress indicators of heart rate variability and electrodermal activity with application in healthcare simulation research. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*, 15(1):39–45, February 2020.

[33] Hugo F. Posada-Quintero and Ki H. Chon. Innovations in electrodermal activity data collection and signal processing: A systematic review. *Sensors*, 20(2), 2020.

[34] Ravil I. Mukhamediev, Yelena Popova, Yan Kuchin, Elena Zaitseva, Almas Kalimoldayev, Adilkhan Symagulov, Vitaly Levashenko, Farida Abdoldina, Viktors Gopejenko, Kirill Yakunin, Elena Muhamedijeva, and Marina Yelis.

Review of artificial intelligence and machine learning technologies: Classification, restrictions, opportunities and challenges. *Mathematics*, 10(15), 2022.

[35] Maximilian Pichler and Florian Hartig. Machine learning and deep learning—a review for ecologists. *Methods in Ecology and Evolution*, 14(4):994–1016, 2023.

[36] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[37] Ronald J Tallarida, Rodney B Murray, Ronald J Tallarida, and Rodney B Murray. Chi-square test. *Manual of pharmacologic calculations: with computer programs*, pages 140–142, 1987.

[38] Gloria Cosoli, Angelica Poli, Lorenzo Scalise, and Susanna Spinsante. Measurement of multimodal physiological signals for stimulation detection by wearable devices. *Measurement*, 184:109966, 2021.

[39] Shahadat Uddin, Asifullah Khan, Muhammad Hossain, and Mohammad Ali Moni. Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, 19(1):281, 2019.

[40] Farhad Maleki, Nikesh Muthukrishnan, Katie Ovens, Caroline Reinhold, and Reza Forghani. Machine learning algorithm validation: From essentials to advanced applications and implications for regulatory certification and deployment. *Neuroimaging Clinics of North America*, 30(4):433–445, 2020. Machine Learning and Other Artificial Intelligence Applications.

[41] Minu Treesa Abraham, Neelima Satyam, Revuri Lokesh, Biswajeet Pradhan, and Abdullah Alamri. Factors affecting landslide susceptibility mapping: Assessing the influence of different machine learning approaches, sampling strategies and data splitting. *Land*, 10(9), 2021.

[42] M. Arif, A. Basri, G. Melibari, et al. Classification of anxiety disorders using machine learning methods: A literature review. *Insights Biomed Res*, 4(1):95–110, 2020.

[43] Sandip S. Panesar, Rhett N. D'Souza, Fang-Cheng Yeh, and Juan C. Fernandez-Miranda. Machine learning versus logistic regression methods for 2-year mortality prognostication in a small, heterogeneous glioma database. *World Neurosurgery: X*, 2:100012, 2019.

[44] Lei Cai, Jingyang Gao, and Di Zhao. A review of the application of deep learning in medical image classification and segmentation. *Annals of Translational Medicine*, 8(11), 2020.

[45] J. Wang, H. Zhu, SH. Wang, and et al. A review of deep learning on medical image analysis. *Mobile Networks and Applications*, 26:351–380, 2021.

[46] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy. A review on deep learning in medical image analysis. *International Journal of Multimedia Information Retrieval*, 11:19–38, 2022.

[47] Jaya Gupta, Sunil Pathak, and Gireesh Kumar. Deep learning (cnn) and transfer learning: A review. *Journal of Physics: Conference Series*, 2273(1):012029, may 2022.

[48] Dário Passos and Puneet Mishra. A tutorial on automatic hyperparameter tuning of deep spectral modelling for regression and classification tasks. *Chemometrics and Intelligent Laboratory Systems*, 223:104520, 2022.

[49] Polipireddy Srinivas and Rahul Katarya. hyoptxg: Optuna hyper-parameter optimization framework for predicting cardiovascular disease using xgboost. *Biomedical Signal Processing and Control*, 73:103456, 2022.

[50] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. *CoRR*, abs/1907.10902, 2019.

[51] Scott M. Lundberg, Gabriel Erion, Hugh Chen, and et al. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2:56–67, 2020.

[52] Niaz Chalabianloo, Yekta Said Can, Muhammad Umair, Corina Sas, and Cem Ersoy. Application level performance evaluation of wearable devices for stress classification with explainable ai. *Pervasive and Mobile Computing*, 87:101703, 2022.

[53] Gabriele Rescio, Andrea Manni, Andrea Caroppo, Marianna Ciccarelli, Alessandra Papetti, and Alessandro Leone. Ambient and wearable system for workers' stress evaluation. *Computers in Industry*, 148:103905, 2023.

[54] Hika Barki and Wan-Young Chung. Mental stress detection using a wearable in-ear plethysmography. *Biosensors*, 13(3), 2023.

[55] Sebastian Mach, Pamela Storozynski, Josephine Halama, and Josef F. Krems. Assessing mental workload with wearable devices – reliability and applicability of heart rate and motion measurements. *Applied Ergonomics*, 105:103855, 2022.

[56] Wonju Seo, Namho Kim, Cheolsoo Park, and Sung-Min Park. Deep learning approach for detecting work-related stress using multimodal signals. *IEEE Sensors Journal*, 22(12):11892–11902, 2022.

[57] Waleed Umer. Simultaneous monitoring of physical and mental stress for construction tasks using physiological measures. *Journal of Building Engineering*, 46:103777, 2022.

[58] Niaz Chalabianloo, Yekta Said Can, Muhammad Umair, Corina Sas, and Cem Ersoy. Application level performance evaluation of wearable devices for stress classification with explainable ai. *Pervasive and Mobile Computing*, 87:101703, 2022.

[59] Xinxia Li, Weiwei Zhu, Xiaofan Sui, Aizhi Zhang, Lijie Chi, and Lu Lv. Assessing workplace stress among nurses using heart rate variability analysis with wearable ecg device–a pilot study. *Frontiers in Public Health*, 9, 2022.

[60] Muhammad Ali Fauzi and Bian Yang. Continuous stress detection of hospital staff using smartwatch sensors and classifier ensemble. In *Studies in Health Technology and Informatics*, volume 285, pages 245–250. IOS Press, 2021.

[61] Ruixuan Dai, Chenyang Lu, Linda Yun, Eric Lenze, Michael Avidan, and Thomas Kannampallil. Comparing stress prediction models using smartwatch physiological signals and participant self-reports. *Computer Methods and Programs in Biomedicine*, 208:106207, 2021.

[62] Anusha A. S., Sukumaran P., Sarveswaran V., Surees Kumar S., Shyam A., Tony J. Akl, Preejith S. P., and Mohanasankar Sivaprakasam. Electrodermal activity based pre-surgery stress detection using a wrist wearable. *IEEE Journal of Biomedical and Health Informatics*, 24(1):92–100, 2020.

[63] Yekta Said Can, Heather Iles-Smith, Niaz Chalabianloo, Deniz Ekiz, Javier Fernández-Álvarez, Claudia Repetto, Giuseppe Riva, and Cem Ersoy. How to relax in stressful situations: A smart stress reduction system. *Healthcare*, 8(2), 2020.

[64] Eric E Kaczor, Stephanie Carreiro, Jared Stapp, Benjamin Chapman, and Premananda Indic. Objective measurement of physician stress in the emergency department using a wearable sensor. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, volume 2020, pages 3729–3738. IEEE, 2020.

[65] Kalliopi Kyriakou, Bernd Resch, Günther Sagl, Andreas Petutschnig, Christian Werner, David Niederseer, Michael Liedlgruber, Frank H. Wilhelm, Tess Osborne, and Jessica Pykett. Detecting moments of stress from measurements of wearable physiological sensors. *Sensors*, 19(17), 2019.

[66] Franci Suni Lopez, Nelly Condori-Fernandez, and Alejandro Catala. Towards real-time automatic stress detection for office workplaces. In Juan Antonio Lossio-Ventura, Denisse Muñante, and Hugo Alatrista-Salas, editors,

*Information Management and Big Data*, pages 273–288, Cham, 2019. Springer International Publishing.

[67] Empatica Website. E4 wristband. `https://e4.empatica.com/e4-wristband`, 2020.

[68] Merna Attia, Fatma A. Ibrahim, Mohamed Abd-Elfatah Elsady, Mohamed Khaled Khorkhash, Marwa Abdelazim Rizk, Jaffer Shah, and Samar A. Amer. Cognitive, emotional, physical, and behavioral stress-related symptoms and coping strategies among university students during the third wave of covid-19 pandemic. *Frontiers in Psychiatry*, 13, 2022.

[69] Alessandro Leone, Gabriele Rescio, Pietro Siciliano, Alessandra Papetti, Agnese Brunzini, and Michele Germani. Multi sensors platform for stress monitoring of workers in smart manufacturing context. In *2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–5, 2020.

[70] Katarina Dedovic, Robert Renwick, Najmeh Khalili Mahani, Veronika Engert, Sonia J Lupien, and Jens C Pruessner. The montreal imaging stress task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain. *Journal of Psychiatry and Neuroscience*, 30(5):319–325, 2005.

[71] Christian Josef Merz, Bianca Hagedorn, and Oliver Tobias Wolf. An oral presentation causes stress and memory impairments. *Psychoneuroendocrinology*, 104:1–6, 2019.

[72] Rossana Castaldo, Luis Montesinos, Paolo Melillo, C James, and Leandro Pecchia. Ultra-short term hrv features as surrogates of short term hrv: A case study on mental stress detection in real life. *BMC medical informatics and decision making*, 19(1):1–13, 2019.

[73] Akbar Dehghani, Omid Sarbishei, Tristan Glatard, and Emad Shihab. A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors. *Sensors*, 19(22), 2019.

[74] Lan lan Chen, Yu Zhao, Peng fei Ye, Jian Zhang, and Jun zhong Zou. Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers. *Expert Systems with Applications*, 85:279–291, 2017.

[75] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 400–408, New York, NY, USA, 2018. Association for Computing Machinery.

[76] Yongbo Liang, Mohamed Elgendi, Zhencheng Chen, and Rabab Ward. An optimal filter for short photoplethysmogram signals. *Scientific data*, 5(1):1–12, 2018.

[77] Elisa Mejía-Mejía, John Allen, Karthik Budidha, Chadi El-Hajj, Panicos A. Kyriacou, and Peter H. Charlton. 4 - photoplethysmography signal processing and synthesis. In John Allen and Panicos Kyriacou, editors, *Photoplethysmography*, pages 69–146. Academic Press, 2022.

[78] Muhammad Zubair and Changwoo Yoon. Multilevel mental stress detection using ultra-short pulse rate variability series. *Biomedical Signal Processing and Control*, 57:101736, 2020.

[79] Varun Chandra, Ankit Priyarup, and Divyashikha Sethia. Comparative study of physiological signals from empatica e4 wristband for stress classification. In *Advances in Computing and Data Sciences: 5th International Conference, ICACDS 2021, Nashik, India, April 23–24, 2021, Revised Selected Papers, Part II 5*, pages 218–229. Springer, 2021.

[80] Mohsen Nabian, Yu Yin, Jolie Wormwood, Karen S Quigley, Lisa F Barrett, and Sarah Ostadabbas. An open-source feature extraction tool for the analysis of peripheral physiological data. *IEEE journal of translational engineering in health and medicine*, 6:1–11, 2018.

[81] Marcus Vollmer. A robust, simple and reliable measure of heart rate variability using relative rr intervals. In *2015 Computing in Cardiology Conference (CinC)*, pages 609–612. IEEE, 2015.

[82] Muhammad Amin, Khalil Ullah, Muhammad Asif, Abdul Waheed, Sana Ul Haq, Mahdi Zareei, and R. R. Biswal. Ecg-based driver's stress detection using deep transfer learning and fuzzy logic approaches. *IEEE Access*, 10:29788–29809, 2022.

[83] Lili Zhu, Petros Spachos, and Stefano Gregori. Multimodal physiological signals and machine learning for stress detection by wearable devices. In *2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–6. IEEE, 2022.

[84] Sevinç İlhan Omurca and Ekin Ekinci. An alternative evaluation of post traumatic stress disorder with machine learning methods. In *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, pages 1–7, 2015.

[85] Siriporn Sawangarreerak and Putthiporn Thanathamathee. Random forest with sampling techniques for handling imbalanced prediction of university student depression. *Information*, 11(11), 2020.

[86] Mohammed Khalilia, Swagata Chakraborty, and Mihail Popescu. Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, 11:51, 2011.

[87] Wangshu Zhang, Feng Zeng, Xuebing Wu, Xuegong Zhang, and Rui Jiang. A comparative study of ensemble learning approaches in the classification of breast cancer metastasis. In *2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*, pages 242–245, 2009.

[88] Pramod Bobade and M Vani. Stress detection with machine learning and deep learning using multimodal physiological data. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 51–57. IEEE, 2020.

[89] Souvik Ghosh, Sumitra Mukhopadhyay, and Rajarshi Gupta. A new physiology-based objective mental stress detection technique with reduced feature set and class imbalanced dataset management. In *2021 IEEE International Conference on Technology, Research, and Innovation for Betterment of Society (TRIBES)*, pages 1–6. IEEE, 2021.

[90] Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data classification: Algorithms and applications*, page 37, 2014.