NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

PEDRO MENDES RODRIGUES

Bachelor Degree in Biomedical Engineering Sciences

AUTONOMOUS ASSESSMENT OF VIDEOGAME DIFFICULTY USING PHYSIOLOGICAL SIGNALS

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
September, 2022

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

AUTONOMOUS ASSESSMENT OF VIDEOGAME DIFFICULTY USING PHYSIOLOGICAL SIGNALS

PEDRO MENDES RODRIGUES

Bachelor Degree in Biomedical Engineering Sciences

Adviser: Prof. Dr. Phil Lopes
Assistant Professor, Lusófona University Lisbon

Co-adviser: Prof. Dra. Maria Micaela Leal da Fonseca
Associate Professor, NOVA University Lisbon

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
September, 2022

Autonomous Assessment of Videogame Difficulty Using Physiological Signals

To my parents and siblings.

Acknowledgements

There were various moments during the accomplishment of my dissertation when, due to hardships and setbacks, I deviated from the path and lost myself. Yet, in those moments of solitude and melancholy, there was always someone willing to reach out and help me get back on track. Without the support, advice and encouragement of these people, this work would not have come to fruition.

To my supervisors Phil Lopes and Micaela Fonseca, a deep gratitude for all the guidance and availability they showed during the realization of this dissertation, and whose opinions and criticisms were fundamental throughout this work.

To my colleagues at HEI-lab, thank you for the precious help you gave me during the design and testing of the game. In particular, to Filipe Pinto for the tremendous patience he has shown and the incredible work he has done with the Wandering Druid.

To my parents and siblings, whom I thank for all the support and affection shown during my academic journey.

To my hometown friends, thank you for always being there, even during this misanthropic phase of my life.

Finally, to all my dearest friends who participated in my experiment, thank you for your availability and contribution that allowed me to build the dataset for this study.

"Experience is the teacher of all things." (Julius Caesar)

Abstract

Given the well-explored relation between challenge and involvement in a task, (e.g., as described in Csikszentmihalyi's theory of flow), it could be argued that the presence of challenge in videogames is a core element that shapes player experiences and should, therefore, be matched to the player's skills and attitude towards the game. However, handling videogame difficulty, is a challenging problem in game design, as too easy a task can lead to boredom and too hard can lead to frustration. Thus, by exploring the relationship between difficulty and emotion, the current work intends to propose an artificial intelligence model that autonomously predicts difficulty according to the set of emotions elicited in the player. To test the validity of this approach, we developed a simple puzzle-based Virtual Reality (VR) videogame, based on the Trail Making Test (TMT), and whose objective was to elicit different emotions according to three levels of difficulty. A study was carried out in which physiological responses as well as player self-reports were collected during gameplay. Statistical analysis of the self-reports showed that different levels of experience with either VR or videogames didn't have a measurable impact on how players performed during the three levels. Additionally, the self-assessed emotional ratings indicated that playing the game at different difficulty levels gave rise to different emotional states. Next, classification using a Support Vector Machine (SVM) was performed to verify if it was possible to detect difficulty considering the physiological responses associated with the elicited emotions. Results report an overall F1-score of 68% in detecting the three levels of difficulty, which verifies the effectiveness of the adopted methodology and encourages further research with a larger dataset.

Keywords: Affective Computing, Emotion Assessment, Physiological Signals, Virtual Reality, Videogames

Resumo

Dada a relação bem explorada entre desafio e envolvimento numa tarefa (p. ex., conforme descrito na teoria do fluxo de Csikszentmihalyi), pode-se argumentar que a presença de desafio em videojogos é um elemento central que molda a experiência do jogador e deve, portanto, ser compatível com as habilidades e a atitude que jogador exibe perante o jogo. No entanto, saber como lidar com a dificuldade de um videojogo é um problema desafiante no design de jogos, pois uma tarefa muito fácil pode gerar tédio e muito difícil pode levar à frustração. Assim, ao explorar a relação entre dificuldade e emoção, o presente trabalho pretende propor um modelo de inteligência artificial que preveja de forma autônoma a dificuldade de acordo com o conjunto de emoções elicitadas no jogador. Para testar a validade desta abordagem, desenvolveu-se um jogo de puzzle em Realidade Virtual (RV), baseado no *Trail Making Test* (TMT), e cujo objetivo era elicitar diferentes emoções tendo em conta três níveis de dificuldade. Foi realizado um estudo no qual se recolheram as respostas fisiológicas, juntamente com os autorrelatos dos jogadores, durante o jogo. A análise estatística dos autorelatos mostrou que diferentes níveis de experiência com RV ou videojogos não tiveram um impacto mensurável no desempenho dos jogadores durante os três níveis. Além disso, as respostas emocionais auto-avaliadas indicaram que jogar o jogo em diferentes níveis de dificuldade deu origem a diferentes estados emocionais. Em seguida, foi realizada a classificação por intermédio de uma Máquina de Vetores de Suporte (SVM) para verificar se era possível detectar dificuldade, considerando as respostas fisiológicas associadas às emoções elicitadas. Os resultados relatam um F1-score geral de 68% na detecção dos três níveis de dificuldade, o que verifica a eficácia da metodologia adotada e incentiva novas pesquisas com um conjunto de dados maior.

Palavras-chave: Computação Afetiva, Avaliação Emocional, Sinais Fisiológicos, Realidade Virtual, Videojogos

Contents

List of Figures

List of Tables

Acronyms

AC	Affective Computing
ANS	Autonomous Nervous System
BVP	Blood Volume Pulse
CV	Cross-Validation
DDA	Dynamic Difficulty Adjustment
DT	Decision Tree
ECG	Electrocardiography
EDA	Electrodermal Activity
EEG	Electroencephalography
EMG	Electromyography
EOG	Electrooculography
GB	Gradient Boosting
GPR	Gaussian Process Regression
GSR	Galvanic Skin Response
HCI	Human-Computer Interaction
HF	High Frequency
HMD	Head Mounted Display
HR	Heart Rate
HRV	Heart Rate Variability
IBI	Inter-Beat Interval

KNN K-Nearest Neighbors

LDA Linear Discriminant Analysis
LF Low Frequency

ML Machine Learning

NB Naïve Bayes
NPC Non-Playable Character

OAA One-Against-All
OAO One-Against-One

PPG Photoplethysmography

RBF Radial Basis Function
RF Random Forest
RMSE Root Mean Square Error

SAM Self-Assessment Manikin
SC Skin Conductance
SCL Skin Conductance Level
SCR Skin Conductance Response
SMOTE Synthetic Minority Over-sampling Technique
SVM Support Vector Machine

TMT Trail Making Test

VAD Valence-Arousal-Dominance
VR Virtual Reality

1 Introduction

1.1 Scope and Context

Emotions are a complex phenomenon that influence many aspects of our everyday life. They assist in decision making, are heavily related to tendencies to action and are a fundamental part of the communication between humans. In fact, the expression of an individual's emotions can considerably change how others perceive the meaning of their messages [1]. This link between emotion, human perception and cognition has been explored by several researchers in the field of psychology and neuroscience and has led to the emergence of several theoretical models of emotion, two of which stand out from the rest, in terms of popularity. These are the discrete emotion model proposed by Ekman [2] and the two-dimensional valence-arousal model proposed by Russell [3].

In recent years, research on the topic of emotions has also extended to the field of Human-Computer Interaction (HCI). HCI is a multidisciplinary field whose main objectives are to study the interaction between man and computer, and minimize the existing barrier between the two through the development of friendly, agile and clear interfaces [4]. The integration of the two fields is the basis of the research area of Affective Computing (AC), which seeks to design systems capable of perceiving and reacting to the emotions of the user. According to Picard [5], designing computers with affective capabilities will not only allow for more natural, effective, and efficient means of HCI, but will also improve the computers' ability to make decisions. It is believed that the progress towards automatic emotion recognition HCI systems will bring significant value to both scientific research and commercial activities, within fields such as those of psychology [6, 7], robotics [8, 9], education [10] and healthcare [11].

One of the applications that could greatly benefit from emotion recognition in an affective computing framework is videogames. As one of the major goals of videogames is to provide entertaining, immersive and emotional experiences to players, affective computing emerges as a promising area of research to further enhance the player experience. In the context of games, the use of emotion recognition offers a way to assess the player's

level of involvement and enjoyment during gameplay, and provide changes to the game to maintain the player in a certain emotional state [12]. In the literature, several emotional frameworks that model the affective state of players, have been proposed in an attempt to characterize the player experience, some of which are based of Csikszentmihalyi's theory of flow [13]. According to these models, strong involvement in the videogame occurs when the abilities of the player match the difficulty of the tasks. As such, in response to the players emotions and competence, the game would adjust challenge such that it would be neither insufficiently or excessively challenging and thus the most engaging possible [14]. For this, the automatic assessment of players' emotions is a necessary requirement.

Human emotions can be detected from different modalities, for example from video and audio analysis of facial expressions, gestures and speech [7, 15]. However, while these sources allow direct access to an individual's emotions and are easy to interpret, they are liable to be manipulated [16], which can compromise the effectiveness of the computer's interpretation. A more reliable source of emotional information is physiological signals [16]. The fact that their manifestation is determined by the Autonomous Nervous System (ANS), whose behavior can hardly be controlled by the subject's will, allows for more reliable and objective results, compared to other sources.

In the literature it is possible to find several experimental designs used for the study of emotions in videogames. These differ mainly in terms of targeted emotions, elicitation methods, data sources or modalities, and classification techniques. In fact, the design of the experimental protocol and associated methodologies should be regarded as a fundamental procedure to the success of the collection and labeling of emotional experiences. This dissertation aims to explore the effects of a videogame's difficulty on player's emotions, based on an experimental procedure that involves the design of a Virtual Reality (VR) videogame, and the acquisition of physiological data and players self-reports. The usage of a Support Vector Machine (SVM) to classify 3 classes of game difficulty, based on the data collected, will be done towards evaluating the validity and usefulness of the proposed game, as a medium for studying the effects of difficulty on the emotions of the players.

1.2 Objectives

The objectives of the present work are related to the major purpose of proposing an automatic system for emotion recognition and difficulty adaptation in a VR videogame, namely:

- Designing an effective emotion elicitation protocol based on a VR videogame;

- Collecting emotional and experience self-reports and corresponding physiological response from three signals (i.e., cardiovascular, eletrodermal and respiratory activities);
- Implement a workflow for the pre-processing and segmentation of the acquired physiological signals;
- Extraction of relevant features from the processed physiological signals;
- Automatic assessment of the difficulty experienced by players during gameplay of the proposed VR game;
- Evaluation of the validity and usefulness of the proposed VR game as a medium to study the effect of difficulty on players' emotions.

1.3 Thesis Outline

This dissertation is divided into six chapters. The first has already been presented, along with the contextualization of the problem and definition of objectives. Chapter 2 comprises the theoretical framework of this study, through the exploration of concepts related to AC, emotion theory, sources of emotional expression, and the theory of SVM classifiers. Chapter 3 consists of a literature review on the main aspects of emotion recognition experiments, where videogames and physiological data are used. Chapter 4 presents a detailed description of the methodologies adopted for the development of the VR game proposed for this dissertation, as well as the definition of the workflow used for treatment of the data (i.e., signal preprocessing, segmentation, feature extraction and classification). In Chapter 5, the most relevant results will be presented and discussed, taking into account the objectives outlined for the dissertation, namely the evaluation of the games' accessibility and ability to elicit different emotions, along with the classification of difficulty taking into account the emotions evoked during gameplay. Finally, Chapter 6 presents the main conclusions of the work based on the previous chapter, and provides some suggestions for improvements and future work.

2

Theoretical Concepts

2.1 Affective Computing

Emotions are a core concept of human intelligence. They enrich human communication, through the use of verbal (i.e., emotional vocabulary) and non-verbal cues (i.e., tone of voice, facial expressions and body postures). But also play a key role on how humans think and behave. In his book, D. Goleman [17] emphasizes on how emotions can compel humans to act, triggering action readiness, and how they influence the decisions they make throughout everyday life. The author stresses that *"emotional intelligence"*, i.e., the ability to understand, communicate and manage the emotions of oneself, along with understanding and showing empathy towards the emotions of others, should be regarded as fundamental part of any natural communication involving humans. In spite of that, most contemporary HCI systems are deficient in interpreting and understanding emotional information and completely lack *"emotional intelligence"*. Human interactions are not limited to interpersonal conversation, but also extend to their surroundings, including computers. Nowadays, it is normal for people to spend more time of there day interacting with a computer than face-to-face with other humans. As the interaction and collaboration between man and machine increases in a variety of environments, it is important to design computer systems that excel towards user-friendliness. One way is through intelligent HCI systems, that are capable of both understanding and adapting towards the individual needs of the user. Users should be able to express their intentions in a natural way, i.e., through verbal and nonverbal communication such as emotions, gestures, and facial expressions. According to Reeves et al. [18] people treat computers the same way as they treat people. Thus, computerized automatic recognition of the users emotional state can therefore be considered a crucial step towards the development of more efficient, flexible and advanced HCI systems.

The desire to integrate emotions into the HCI loop, has led to the emerging of the research field called affective computing AC. The term AC was first introduced in 1995 as a type of HCI whose goal, amongst others, is to *"give computers the ability to recognise*,

express, and in some cases, 'have' emotions" [5]. Since then, AC has been at the core of a multitude of proposed emotionally intelligent systems for recognition, interpretation, visualisation and stimulation of affective experiences [19]. Mathias et al. [20] summarize the three procedures, figure 2.1, which substantiate the design of traditional affective computing systems as follows: emotion recognition, the emotional behavior generation and emotion elicitation. Emotion recognition consists of an affect detection system that recognizes whether the user is experiencing positive or negative emotions. This can be achieved by monitoring different cues, which include both verbal and nonverbal modalities, for instance speech, facial expressions, gestures or physiological responses [19]. Following detection, the built-in reasoning and action selection mechanism will choose the optimum system response and adapt itself, which will evoke a reaction in the user. These three phases model the iterative process defined as affective loop [8, 20].

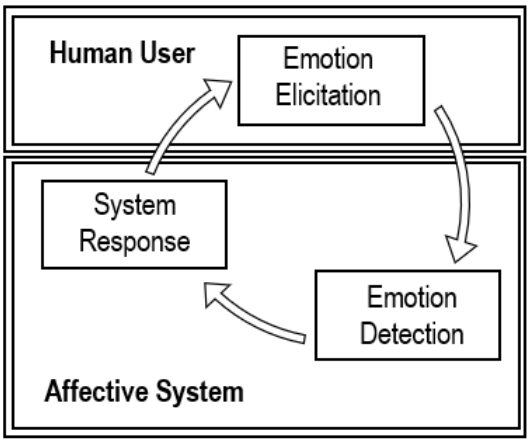

Figure 2.1: The Affective loop which encompasses 3 steps: emotion detection, emotional behavior generation (or system response) and emotion elicitation. Image adapted from [20].

Modern AC systems encompass a wide variety of applications, which range from robotics [8, 9], to education [10] and healthcare [11]. This dissertation is focused on its application to videogames, known as affective gaming [21]. Affective videogames are capable of recognising emotions during gameplay and dynamically adjust specific game features accordingly. Through dynamic emotionally-driven adaptation, videogames have the ability of inducing felicitous emotional states, promoting engagement, while straining away from others such as boredom or anxiety [12, 22]. The application of AC to videogames is further detailed in chapter 3.

2.2 Emotion Theory

As stated, emotions are a complex phenomenon that influence many aspects of our everyday life: they play a key role in decision making, are heavily related to tendencies to

action and are actively present during our social interactions. Moreover, in order for the affective loop to implement adjustments and elicit the desired emotions, it requires an emotional framework that provides an adequate interpretation of the user psychophysiological state and substantiates the system response. Nonetheless, there is no common framework that can be used to answer the question of what an emotion is. Hence, in order to better understand emotions, several theoretical models of emotions have been proposed, each classifying emotions through different representations [23]. The following section will focus on two models: the discrete emotional model, figure 2.2, as proposed by Ekman [2], and the two dimensional valence-arousal model, figure 2.3, initially proposed by Russel [3].

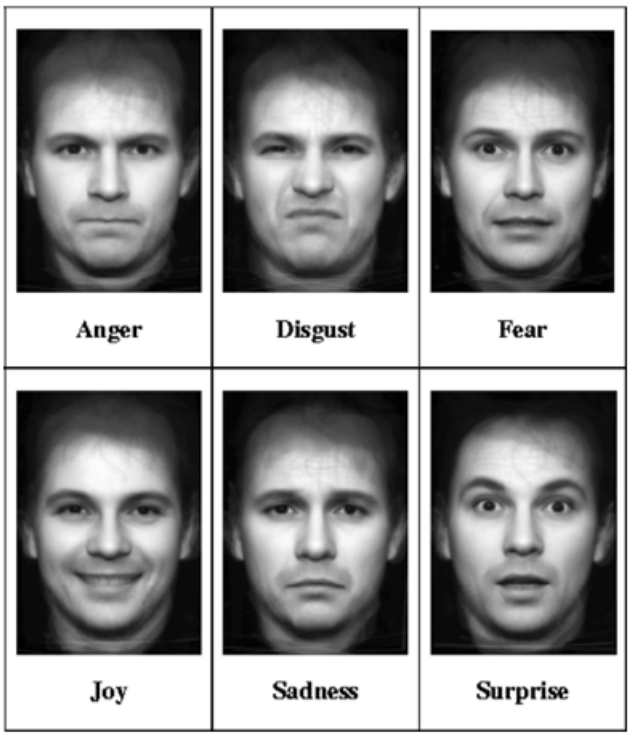

Figure 2.2: Ekman's "Big Six"emotions [24]: "joy", "sadness", "fear", "anger", "disgust" and "surprise".

Discrete models divide emotions into a set of basic emotions. These are regarded as being biologically distinctive and innate emotional responses, whose expression is fundamentally the same for all individuals [2]. However, this theory lacks reliable scientific evidence that substantiates this assumptions [25]. According to this models, an emotion is considered basic if it is "elementary", meaning it cannot be decomposed into a combination of other emotions, but can be combined with other basic emotions to construct several non-basic emotions [26]. Although it seems advantageous, as it enables to virtually construct any emotion as a combination of others, there is a lack of consensus among researchers on the number and set of emotions, as well as in the way they are combined

[26]. According to Picard [5], researchers have proposed more than twenty discrete emotional sets, however, despite the diversity in categories of emotions, fear, anger, sadness, and joy are the most commonly listed.

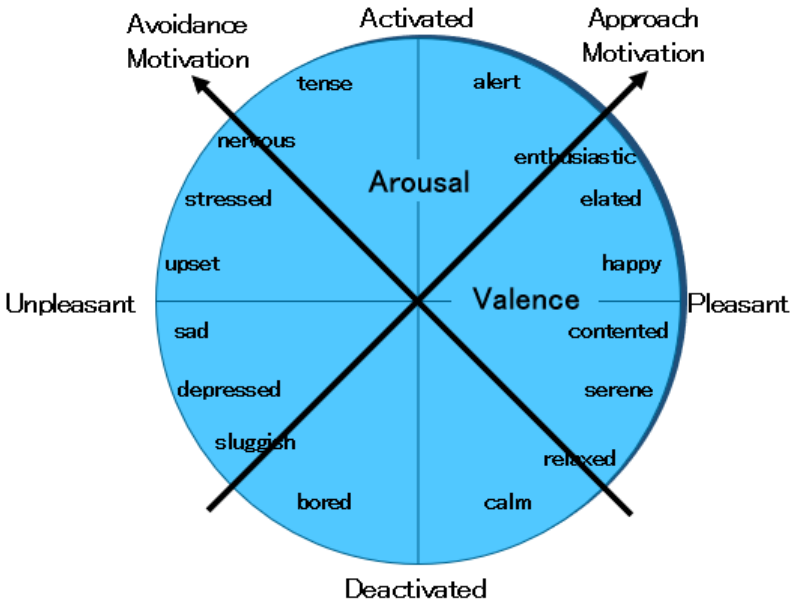

Figure 2.3: Russel's Emotional Circumplex Model[27]. The core bipolar bi-dimensional space for representation of emotional states, determined by valence and arousal. Quadrant I (positive valence and arousal): alert, enthusiastic, elated, happy; Quadrant II (negative valence, positive arousal): tense, nervous, stressed, upset; Quadrant III (negative valence, negative arousal): sad, depressed, sluggish, bored; Quadrant IV (positive valence, negative arousal): calm, relaxed, serene, contented. The tendencies to approach or avoidance provide alternative axes to the 2D affect space.

In contrast to the discrete model, dimensional models organise emotional states using two or more dimensions, meaning emotions are represented within a multi-dimensional continuous space. The circumplex model of affect [3], which is commonly used in affective studies, defines a two-dimensional independent bipolar space, in which an emotion is represented by the level of pleasure (or valence) *(x-axis)* and arousal *(y-axis)*. While valence denotes the bipolar subjective evaluation of the emotional feeling (i.e., pleasantness), arousal expresses the bipolar level of mental activation or engagement. Mendl et al. [27] adapted the design to account for the idea that most living organisms possess basic approach/avoidance responses that serve to guide behaviour towards *"maximizing acquisition of fitness-enhancing rewards and minimizing exposure to fitness-threatening punishers"*. The authors hypothesise that in more complex animals, such as humans, discrete

emotions such as fear or anger result from the association between these core affect processes, and cognitive appraisals of the self and the environment [27]. As represented in figure 2.3, the two functions (axes) underlying this behavioural system lie at 45° to the pleasure and arousal axes, one indicates propensity to move towards a desired stimuli (Approach), while the other indicates propensity to move away from a stimuli or situation (Avoidance).

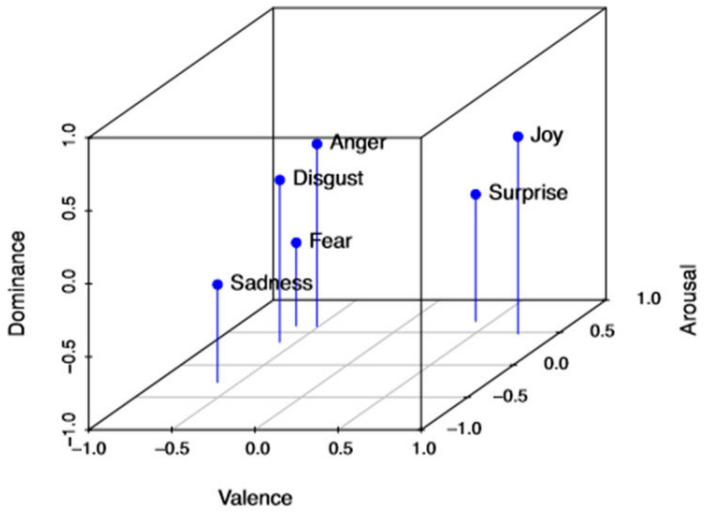

Figure 2.4: The VAD model introduced by Russel [28]. The models encompasses the three dimensions. The values for the six basic emotions are represented, as to provide a correlation between the VAD model and the discrete model for emotions.

Alternatively to the two-dimensional approach, some authors argue that the addition of a third dimension is required for a complete representation of the semantic space of emotion. For instance, Russell et al. [29] introduced the Valence-Arousal-Dominance (VAD) model, a three-dimensional space in which the emotional state is described using valence, arousal and dominance, as represented in figure 2.4. While the first two have similar meanings to the bi-dimensional model, dominance reflects the level of control of the emotional state, which ranges from submissive to dominant. The usefulness of the dimension of dominance becomes clear, for instance to discriminate fear from anger, which in the valence-arousal space share a similar projection. However, the use of a third dimension is generally discarded in most studies, as it represents only a small part of the variance in participant judgments and tends to overlap with arousal ratings [3].

Even though there is no clear consensus in psychology on which of the categories of models better depicts the underlying nature of emotion [30], dimensional models, namely the two-dimensional valence-arousal space, present several advantages over discrete models. For instance, as each emotion is represented by a coordinate system, this implies that any emotion can be represented as a point in this space, even in circumstances where

there is no discrete label associated with the point. Moreover, it is possible to directly map discrete labels onto the dimensional space [3], although there is some variability in the mapping amongst individuals. Another advantage is the possibility to represent emotional intensity [3], since regions near the center of the space are indicative of low intensity and neutral emotions, and regions at the periphery suggest high intensity. Finally, for a computational representation of emotions, dimensional models are less restrictive than discrete categories [31], while also being more adequate for performing calculations [32].

2.3 Multimodal Sources of Emotion

Emotions are a mental state characterized by distinct levels of energy and whose expression is stimulated by changes, on both conscious and non-conscious processes, that combine to form the emotional experience. Social psychologists agree that emotions can be separated into three critical components, each playing a role in the function and purpose of the emotional response [33]:

- **Subjective or cognitive component of emotion** refers to the emotional and cognitive impact derived from human experience, and translates to *"How you experience emotion"*.
- **Physiological component** corresponds to the physiological response a human experiences during an emotion. It can be interpreted as *"How your body reacts to the emotion"*.
- **Behavioural component** encompasses the non-verbal response patterns derived from the emotional experience - *"How you behave in response to the emotion"*.

Emotions are related to changes on multiple variables within each on of these components, which translates into emotional manifestations that can occur both internally and externally. Facial expressions [6], gestures, postures [7] and speech [15] are clear examples of external manifestations which have been explored by several studies in the field of emotion recognition. Nevertheless, these modalities have disadvantages, namely being dependent on culture, gender and age [34]. As discussed in [34], although some core components of emotions are universal and presumably biological, the meaning of emotional expression can change depending on nationality and ethnicity. Furthermore, individuals may hesitate to fully express themselves, when such behaviours are socially undesirable with respect their social environment [16].

Alternatively, emotional information can also be derived from physiological signals, as they also provide patterns that reflect emotional expressions [35]. For instance, an increase in Heart Rate (HR) when faced with feelings of fear, or increased sweating during

periods of high anxiety. This modality provides a continuous measure of affect whose expression is directly correlated with the activity of the ANS. Moreover, as the ANS activity can hardly be controlled by the individual's conscious effort, this allows to overcome the social masking problem, making the emotional assessment through physiological signals more reliable and unbiased [36, 37]. Although less perceptible to the naked eye, physiological cues can be directly evaluated from both the peripheral and central nervous systems, by using several non-intrusive and wearable biosensors [38]. Nonetheless, this modality has disadvantages, starting with interpretability, which is more complex and difficult compared to other modalities [39]. Additionally, most biosensors suffer from high susceptibility to motion artifacts, power line noise and noise derived from poor electrode contact [38, 39]. These disadvantages, if not taken into account, can compromise the effectiveness and accuracy of the emotional assessment.

Although several emotion recognition studies have been successful in detecting emotions using a single physiological modality, (e.g., [40, 41]) the ambiguity and complexity of the physiological response make it difficult to map uniquely physiological patterns onto specific emotional states [39]. For this reason, methods such as fusion of different physiological modalities have been widely used (e.g., [39, 42–44]) to enhance the reliability and efficacy of the emotion recognition system. This strategy can increase emotion recognition accuracy, since different modalities often complement each other, while also promoting robustness [45], by discarding anomalous signal behaviours, that are not caused by emotion elicitation.

2.4 Biophysical Signals

In light of the objectives outlined for the dissertation and taking into account the advantages of using physiological modalities over the alternatives described in the previous section, it was determined that this study will use the fusion of several physiological modalities to evaluate the emotional state of players, as a consequence of the gameplay. Additionally, considering that emotions are mostly expressed by means of internal body manifestations, and that Electrocardiography (ECG), Electrodermal Activity (EDA) and respiratory activities [38, 39], together with electroencephalography (EEG) are some of the modalities most commonly found in emotion assessment studies, these were considered for this study. However, only the first three were included, since the EEG sensor is physically incompatible with the VR headset.

2.4.1 Electrodermal Activity

EDA, also known as Skin Conductance (SC) or Galvanic Skin Response (GSR) [46], is a physiological marker of the human body that is associated with changes in the electrical properties of the skin (i.e., resistance or conductance), due to the secretion of sweat

[47]. There are three types of sweat glands distributed over the entire surface of the skin: eccrine, apocrine and apoeccrine [48]. These are exclusively innervated by the sympathetic nervous system [48]; and amidst the three, the eccrine sweat glands hold the most value in terms of affective information, as they are innervated by the sympathetic nerves (primarily the sudomotor nerves [47]) that accompany various psychological processes, including emotional arousal [46, 49]. This makes EDA an ideal measure for sympathetic activation and an adequate biomarker for physiological arousal [50].

EDA measurements can be performed by applying an electrical potential between two points in the skin and measuring the flow of current between them [50]. The magnitude of the recorded signal will depend on the density of sweat glands, their relative size and the output of individual glands, which varies between individuals [48]. As such, although eccrine sweat glands are distributed across the whole body, EDA is preferably measured at the palmar sites of the hands or the feet, since it's where the density of eccrine sweat glands is the highest [48]. The EDA signal can be further decomposed into two quantitative components for a more detailed analysis of the emotional response, these are the tonic component and the phasic component [47, 51].

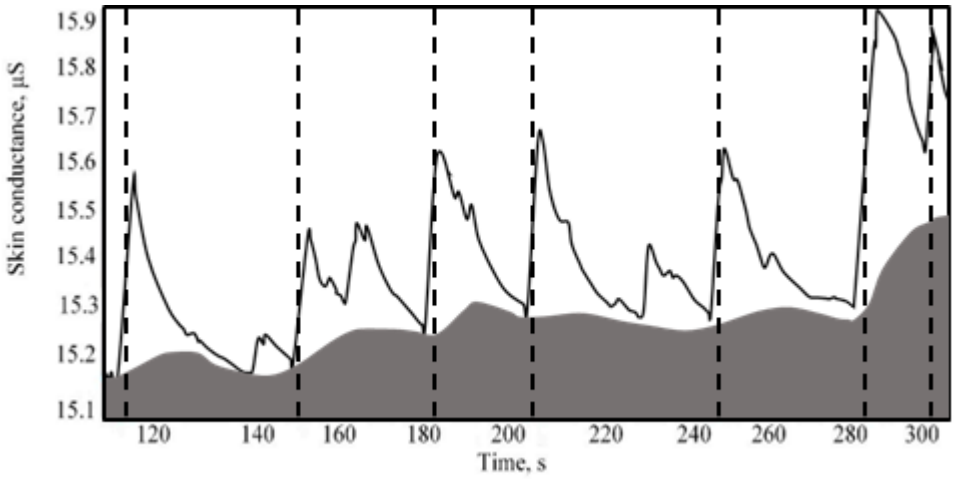

Figure 2.5: Expected behaviour of an EDA signal [38]. The grey area evidences the tonic component of the signal. The white area evidences the phasic component. The dashed line indicates the moment of delivery of the stimulus. Note that not all SCR peaks are preceded by a trigger.

The tonic component comprehends the slower acting elements (i.e., measured in minutes) and background characteristics of the signal that aren't associated with the onset of a specific stimulus. The most common measure of this component is the Skin Conductance Level (SCL), which has been shown as a sensitive and valid indicator of arousal [51]. The other component, the phasic component, refers to the faster changing elements

of the signal (i.e., measured in seconds) - Skin Conductance Response (SCR) - which are distinguished as either event-related or non-specific.

The Event-Related Skin Conductance Response (ER-SCR) is the component of the SCR associated with specific and identifiable stimulus responsible the momentary activation of the sympathetic branch [51]. Non-Specific Skin Conductance Responses (NS-SCR), on the other hand, are spontaneous variations in the SCR which do not occur as a result of a given stimulus or artifact. NS-SCRs can also provide relevant information about the level of stress, anxiety and cognitive load [50]. By looking to figure 2.5, the reader can have a better understanding of the behaviour intrinsic to the two components.

Despite its high popularity in the field of emotion recognition, EDA has some disadvantages, such as high sensitivity to motion artifacts and other environmental factors such as temperature and humidity [52, 53]. However, the main drawback corresponds to the lack of information related to the valence dimension. Even though EDA is a good indicator of arousal level, helping to differentiate between intense and relaxing emotions, it does not allow for an accurate discrimination between positive and negative emotions [38]. This limitation is usually offset by the additional implementation of other emotion recognition methods, capable of providing more conclusive information about the valence level.

2.4.2 Cardiovascular Activity

A healthy heart is characterized by a dynamic relationship between the sympathetic and parasympathetic branches of the ANS [54]. This relationship ensures adequate extrinsic regulation over the intrinsic cardiac system in response to the body's needs. The parasympathetic branch is predominantly active during periods of low stimulation, (metaphorically referred to as the "rest and digest" state [54]), and is responsible for suppressing cardiovascular activity, mainly by decreasing the frequency of the cycle of relaxation and contraction of the heart [54]. In contrast, cardiac sympathetic excitation promotes cardiovascular activity, by increasing the rhythm and the contractile strength of the cardiac muscle ("fight or flight" response [54]). Nevertheless, although antagonistic, the complex relationship between the sympathetic and parasympathetic branches should not be described as a "zero-sum"system (i.e., a decrease in activity in one of the branches does not necessarily lead to an increase in activity in the other branch and vice-versa) [55]. Cardiac changes caused by the ANS can be detected using an ECG, which is a conventional diagnostic method that measures the electrical and cardiovascular activity of the heart [56]. Typically, ECG recording procedures are performed with electrodes that can vary in number and placement [38, 56], hence the choice of a suitable configuration depends on the type of analysis performed on the acquired signal [56].

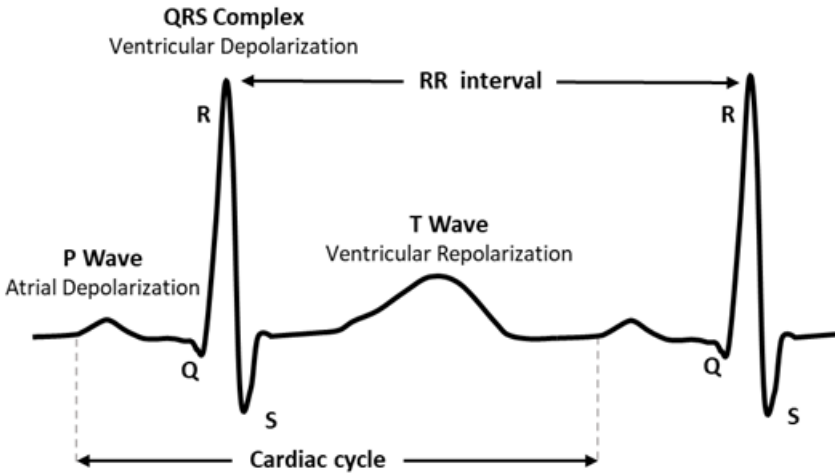

Figure 2.6: ECG waveform [56]. The cardiac cycle and the three electrophysiological events materialized by P wave, QRS complex and T wave. The distance between R peaks is where the IBI is usually calculated for HR detection.

As shown in figure 2.6, in a normal ECG signal there are three-segmented waves visible in each cardiac cycle. These are the P wave, which materializes from the depolarization of the atrial muscle, followed by the QRS complex, evidence of the depolarization of the ventricular muscle, and the T wave, which appears a few milliseconds after the QRS complex and is evidence of the ventricular repolarization [56]. From the ECG, it is possible to derive the HR, from estimation of the Inter-Beat Interval (IBI) between consecutive R peaks; moreover, knowledge of the IBI and/or HR allows for the calculation of the Heart Rate Variability (HRV), which corresponds to the variation in the time interval between consecutive R-R intervals [55, 56].

Due to the physiological interrelationship between the heart and the brain, ECGs are one of the most used physiological signals for emotion recognition, given their quality and rich information on human emotions [54–56]. Thus, in the field of emotion recognition, both HR and HRV have been proven has legitimate indices for physiological assessment of various emotional states. For instance, HR is informative in the sense that a low HR evidences a state of relaxation, and an increased HR is indicative of frustration/mental stress [38]. Similarly, Thayer et al. [57] verified a correlation between high HRV and increased emotional self-control, and in [58] it is observed that an increase in the HRV is accompanied by a decrease in anxiety.

2.4.3 Respiratory Activity

Breathing involves a set of complex interactions between the central nervous system, respiratory-related motor neurons, and respiratory muscles [59]. The motor commands responsible for the intermittent contraction of the inspiratory and expiratory muscles

originate from sophisticated neuronal networks in the brain, which modulate the respiratory rhythm and tidal activity [59]. Although breathing is primarily regulated in response to the metabolic necessities of the body, it also co-exists with emotions, meaning that the respiratory motor output is not only controlled to ensure the body homeostasis, but also constantly responds to changes in emotions, including sadness, happiness, anxiety and fear [20]. This makes the respiration signal (figure 2.7) an encouraging intermediary for emotion assessment, as it has been observed that respiratory parameters such as velocity and depth are indicative of arousal and valence [20, 38, 60]. For instance deep and slow breathing often reflects a relaxed state, while deep and fast breathing shows excitement that is bolstered by happiness, anger, or fear, and rapid shallow breathing suggests tense anticipation (i.e., concentration or fear) [60], whereas shallow and slow breathing evinces states of withdrawal, depression or calm happiness [20]. Nonetheless, this patterns are highly user-dependent; Masaoka et al.[59] emphasize that differences in the respiratory patterns, namely frequency, during mental stress and physical load are related to individual traits.

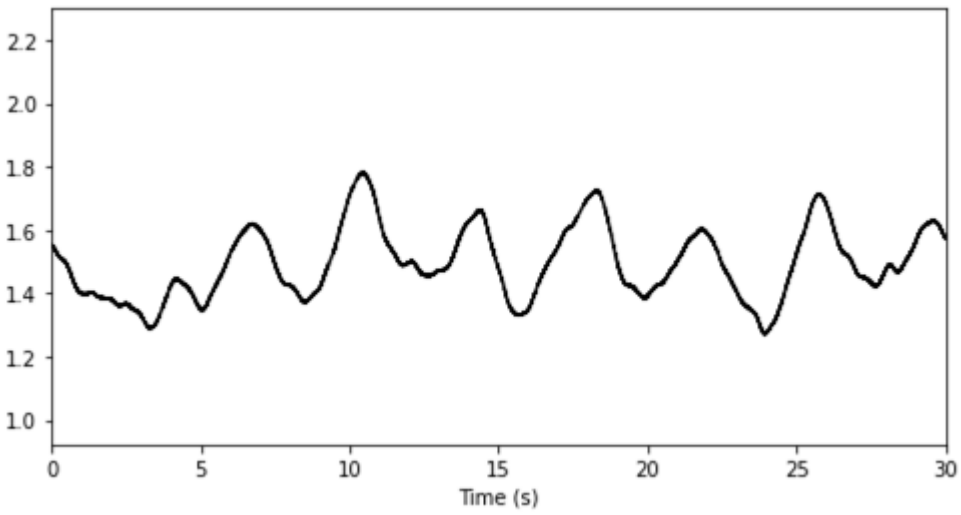

Figure 2.7: Respiratory signal. Typical unfiltered waveform of a respiration signal (measured at rest through a piezoelectric sensor band).

Measurement of the respiratory activity can be achieved directly through several methods, of which the most common relies on the use of resistive wire strains placed around the chest or at the diaphragm level [61]. Moreover, it is also possible to extract respiration features indirectly through analysis of ECG or Photoplethysmography (PPG) measurements [61]. A more detailed review of the various methods used for respiration measurement and processing is provided in [61].

Despite having been appointed as a valid biosignal in the field of emotion recognition, measurement of the respiratory activity for psychophysiological assessment is less common in comparison with EDA or ECG. As referred in [38], the nature of the signal and its' high sensitivity to various external factors, such as fatigue and body movement artifacts, limits the application of this modality. As a result, evaluation of the respiration is often used as a complimentary method along with other emotion recognition and evaluation methods [38].

2.5 Support Vector Machine (SVM)

SVM belong to a set of supervised machine learning algorithms used for solving classification and regression analysis problems. These algorithms which were originally introduced by V. Vapnik and C. Cortes, and published in 1995 [62], have since been successfully applied in several areas, including image-based analysis and classification, computation biology, medical decision support [63] and finance [64]. In particular, this set of supervised learning methods has seen extensive applications in emotion recognition [39]. Their popularity is related to the simplicity of the model for obtaining good generalization solutions, through analytical balancing of the trade-off between decision rule complexity and error frequency [62].

In the context of classification problems, SVMs aim to find the separation function, also called separating hyperplane, that produces the optimal separation of classes [62]. The found hyperplane function, which is determined by training the SVM using subsets of data, intends to maximize the margin between the training patterns and the decision boundary, while simultaneously allowing for a good generalization ability of the model [65]. Moreover, the SVM algorithm finds the best hyperplane, by considering the data points that are closest to it (i.e., support vectors), meaning it doesn't use all the points to specify the hyperplane that maximizes the margin between them. This property is said to make SVMs less affected by the dimension of the feature space [66]. In the following section, an analysis of the mathematical formalism of linear and non-linear SVMs, for cases of binary classification will be made. Subsequently, we will address the classification cases with more than two classes and some concrete aspects of its implementation in the context of the present work.

2.5.1 Linear SVMs

Linear SVMs are algorithms that allow for the definition of linear boundaries between points of a dataset, separating them into two classes [65]. Next, it will be discussed how these classification methods are formulated for linearly separable dataset cases and how they can be extended to non-linearly separable datasets.

2.5.1.1 Hard Margin Linear SVMs

Considering a set of n training data points x_i ($i = 1, 2, ..., n$), where $x_i \in \mathbb{R}^n$, and there respective class labels $y_i \in Y$ where $Y = \{-1, 1\}$. The objective of linear SVMs is to find a hyperplane for the training dataset x_i that linearly separates the data and maximizes the margin between the data points belonging to each of the classes y_i, while retaining good generalization ability (i.e., avoiding overfitting) [65]. Thus, points x which lie on the hyperplane satisfy the following equation:

$$w^T x + b = 0, \tag{2.1}$$

where w is a vector of dimension N perpendicular to the hyperplane and b is a bias term. Defining d_+ and d_- as the shortest distance from the separating hyperplane to the closest positive and negative class (i.e., +1 and −1), respectively, the margin of the separating hyperplane, d, will be defined by $d_+ + d_-$ [67]. In this case, the SVM will look for the separating hyperplane with the highest margin, satisfying the following conditions, (figure 2.8):

$$\begin{cases} w^T x_i + b \geq +1 \text{ if } y_i = +1 \\ w^T x_i + b \leq -1 \text{ if } y_i = -1 \end{cases} \tag{2.2}$$

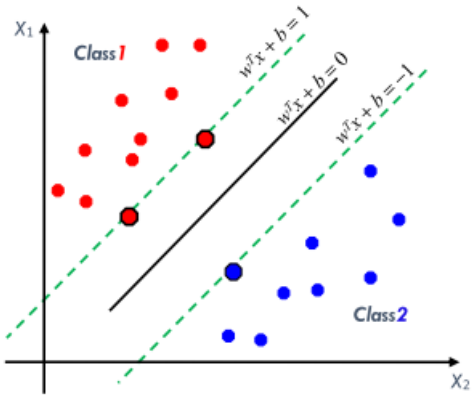

Figure 2.8: Representation of a binary classification using linear SVMs with hard margins [65].

However, there are many margins which can be considered as the boundary of each class, since the generalization region for the hyperplane can be anywhere between 1 and −1 [65]. Thus, the optimal hyperplane will be determined by maximizing the distance d, between the margins, using the following equation:

$$d(w, b; x) = \frac{|(w^T x + b - 1) - (w^T x + b + 1)|}{\|w\|} = \frac{2}{\|w\|} \tag{2.3}$$

Additionally, equation (2.2) can be described by the following inequality:

$$y_i(x_i w^T + b) - 1 \geq 0 \; \forall i \tag{2.4}$$

From equation (2.3), it is verified that the maximization of the distance d can be achieved through minimization of $\|w\|$. Hence, the present minimization problem is named the primal problem [67] and will be formulated as follows :

$$\begin{aligned} min_{w,b} &= \frac{1}{2}\|w\|^2 \\ \text{s.t } &\textit{equation (2.4)} \end{aligned} \tag{2.5}$$

In the case of hard margin SVMs, it is necessary to ensure that there are no training data points between the class margins (linearly separable). This limitation is stipulated by the constraint represented in inequality (2.4). The present problem can be solved using the minimization of the following Lagrangian function that includes the previous restriction in the objective function:

$$L_P(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i (y_i(w^T x_i + b) - 1) \tag{2.6}$$

The minimization of the Lagrangian function will imply the maximization of α_i and the minimization of w and b. Therefore, the resulting stationary (saddle) point will have to fulfill the following Karush–Kuhn–Tucker (KKT) conditions [65]:

$$\frac{\partial L}{\partial w} = 0 \implies w_0 = \sum_{i=1}^{N} \alpha_i x_i y_i \tag{2.7}$$

and

$$\frac{\partial L}{\partial b} = 0 \implies \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{2.8}$$

Substituting equations (2.7) and (2.8) in equation (2.6) yields the Dual Langragian formulation [67], which gives the general equation of the SVM for a a linearly separable case:

$$\begin{aligned} \max \quad & L_D(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ \text{s.t} \quad & \begin{cases} \alpha_i \geq 0 \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \end{cases} \end{aligned} \tag{2.9}$$

Considering the above equation and its constraints, and recalling the second part of equation (2.6), it is possible to conclude that α_i will be greater than 0 if and only if the corresponding input data points are located on the margins. These points are called support vectors and are the only ones that interfere in the determination of the optimal hyperplane equation, since for the remaining points α_i is equal to 0 [67]. Although equations (2.6) and (2.9) arise from the same objective function, the constraints applied to both formulations are different, thus each one was associated with a different Lagrangian label (P for primal, D for dual). So the solution is found by either minimizing L_P or maximizing L_D.

By solving equation (2.9) it is possible to find the support vectors and their corresponding data. Hence, the value of w is obtained using equation (2.7), and the parameter b is calculated from the average of the values of b for all support vectors. The following equation presents the formula for calculating b, where N_S is the number of support vectors and S is the support vector set[65].

$$b_0 = \frac{1}{N_S} \sum_{S=1}^{N_S} (y_S - w^T x_S) \tag{2.10}$$

2.5.1.2 Soft Margin Linear SVMs

Even in cases where the data is not linearly separable, i.e. cases where the constraints presented above are not satisfied, it may be possible to apply linear SVMs effectively. Thus, Soft Margin SVMs are an extension of Hard Margin SVMs, with the introduction of a penalty function, which measures the distance (ξ) between a point and the margin of the respective assigned class [65, 67], in cases of misclassification, figure 2.9. Hence, the penalty function can be defined as:

$$F(\xi) = \sum_{i=1}^{N} \xi_i, \quad \text{for } \xi_i \geq 0 \tag{2.11}$$

Considering the addition of the soft margin penalty, for a linearly non separable case, the optimization problem (2.5) is reformulated as follows:

$$\begin{aligned} min_{w,b} &= \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} \xi_i \\ y_i(x_i w^T + b) &\geq 1 - \xi_i \end{aligned} \tag{2.12}$$

where the parameter C is a positive constant that specifies the trade-off between maximizing the margin and minimizing the number of misclassifications. The higher the value chosen for C, the greater is the penalty assigned to margin violations, and consequently,

the margin between classes will be smaller [67]. Contrary to what happened in the hard margin case, the inclusion of the penalty function, allows for some misclassification, i.e., it allows data to appear between the margins (figure 2.9).

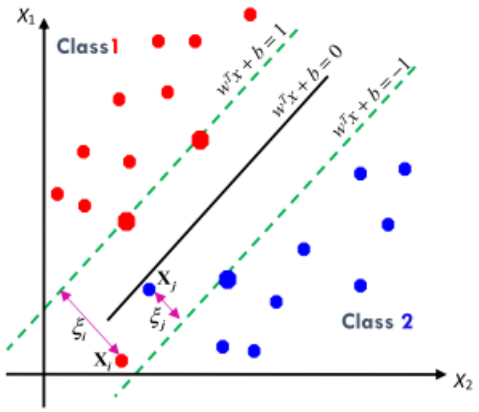

Figure 2.9: Representation of a binary classification using linear SVMs with soft margins [65].

The solution to the optimization problem is found in a similar manner to the previous one, by applying the Lagrangian function and equating its derivatives to zero. Hence, to find the optimal hyperplane, the new problem can be formulated as follows:

$$\begin{aligned} \max \quad & L_D(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ \text{s.t} \quad & \begin{cases} 0 \leq \alpha_i \geq C \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \end{cases} \end{aligned} \tag{2.13}$$

As observed above, the formulation of the problem for soft margin linear SVMs is very similar to that of the case of the hard margin linear SVMs, as it only differs in the constraint for α_i which is forced to be less than or equal to C.

2.5.2 Non-Linear SVMs

Linear SVMs are effective in cases of linearly separable datasets or datasets with an approximately linear distribution. Otherwise, its application is not suitable, since the model will lose its generalization ability, which allows for good results during the test phase [65]. Then, to solve the non-linearity issue, SVMs map the training set from the input space to a higher dimension space known as feature or Hilbert space. The mapping of the input data, x_i, to the feature space is done through a transformation function $\Phi(x)$, which allows the classification of the training set to be done by a linear decision function

[65]. To find the optimal hyperplane, a soft margin linear SVM is applied, which results in a maximization problem identical to that of (2.13).

$$\begin{aligned} \max \quad & L_D(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (\Phi(x_i).\Phi(x_j)) \\ \text{s.t} \quad & \begin{cases} 0 \ \leq \alpha_i \geq C \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \end{cases} \end{aligned} \tag{2.14}$$

As can been seen, the new formulation becomes dependent of the calculation of the inner products $\phi(x_i).\phi(x_j)$. Nevertheless, it is possible to map the original data to a possibly infinite dimensional Hilbert space, making the calculation of the inner products $\phi(x_i).\phi(x_j)$, at the computational level, excessively expensive, for high dimensional feature spaces. Therefore, to make the process faster, the inner product term is replaced by a function called Kernel, $K(x_i, x_j)$. It should, however, be noticed that the kernel trick is valid if and only if for a non-linear vector function, for instance $g(x)$, meets the following integration condition [65]:

$$\int K(x_i, x_j) g(x_i) g(x_j) dx_i dx_j \geq 0 \tag{2.15}$$

There are several Kernel functions which comply with the condition above. Table 2.1 describes those that are available in Python's Scikit-learn package [68]. Note that, the γ parameter, present in all the kernel functions except for the linear, is a constant whose meaning and function varies depending on the type of kernel, however in the case of the Gaussian RBF, which is used in this dissertation, the γ determines the degree of similarity between two points [68]. Thus, when the γ value is low, it means that two points will be considered similar even if they are far apart. Similarly, in cases where the value of γ is high, two points will be considered identical only if they are close to each other. The determination of the value of γ is fundamental for a good performance of the SVM, since, if the parameter γ is too low, the classifier will not capture the complexity of the training data, while, if the value of γ is too high the model will be overly complex and may lose its ability to generalize and, in turn, will underperform in the test phase [68].

Table 2.1: Different Kernel Functions readily available in the Scikit-learn package [68].

Type of Classifier	**Kernel Function**
Linear	$x_i.x_j$
Polynomial of degree ρ	$(\gamma x_i.x_j + r)^\rho$
Gaussian RBF	$exp(-\gamma \|x_i - x_j\|^2), \gamma > 0$
Sigmoid	$tanh(\alpha(x_i.x_j) + \vartheta)$

Knowing that, under this circumstances, the hyperplane is given by:

$$d(x) = w^T \Phi(x) + b, \tag{2.16}$$

Finding the optimal hyperplane, will require the determination of the parameters w and b. However, the use of the kernel trick makes the direct calculation of w unnecessary. Taking into consideration that w is defined in the feature space as:

$$w = \sum_{i=1}^{N} y_i \alpha_i \Phi(x_i), \tag{2.17}$$

By applying the kernel trick and replacing the w on equation (2.10), by the above function, the b parameter can be determined using the following equation:

$$b = y_i - \sum_{i,j=1}^{N_S} y_i \alpha_i K(x_i, x_j), \tag{2.18}$$

Finally, the general equation of the hyperplane is given by substituting the equation(2.17), into equation (2.16), while considering an adequate kernel function [65].

$$d(x) = \sum_{i=1}^{N} y_i \alpha_i K(x, x_i) \tag{2.19}$$

2.5.3 Multiclass SVM

Although SVMs were originally created for binary classification problems, these algorithms can also be used for applications involving multiclass classification. As such, to perform classification tasks on more than two classes with SVMs, this types of problems are decomposed into several binary classification tasks, using one of the following methods: One-Against-All (OAA) or One-Against-One (OAO). Other methods can be employed, however this two are the most common [69, 70].

The OAA approach decomposes the multiclass problem into k binary problems, where each problem compares one of the k classes with the remaining $k-1$ classes [69]. This method encodes the target class as $+1$ and the remaining classes with -1, and classifies unseen data accordingly. Furthermore, this approach requires each model to compute a class membership probability during the classification task, in order to solve the ambiguity generated by cases in which different binary models return $+1$ [71]. In turn, the OAO, like the OAA, is a heuristic method which separates a multiclass classification problem into several binary classification problems, however, unlike the OAA, OAO creates a set of SVM classifiers for all possible pairs of classes [69]. Thus, for k classes, $k(k-1)/2$ classifiers are constructed, each trained only on data from two classes. The classification

process will then produce several class labels, one for each classifier, and the label that occurs the most is assigned to the sample [69]. In this dissertation, the OAO method was chosen, since it provides an adequate balance between accuracy and running time, particularly in situations were the kernel trick is employed [70].

Moreover, regarding the choice of the most appropriate kernel, it was decided to test only two kernel configurations (Linear and Radial Basis Function (RBF)). This decision was based on three criteria, which are kernel popularity, behaviour and complexity. Thus, starting with the linear kernel, they require less training time and resources compared to other kernels, since they work directly in the input space. Additionally, in certain situations, in which the classifier is allowed to make some errors (Soft Margin condition), these kernels have accuracy's comparable to those of non-linear kernels, with the added benefit of providing much faster training and testing speeds [72]. However, in cases where the data cannot be correctly separated by a linear decision function, that is, cases where the relationship between classes and attributes is non-linear, it is necessary to resort to other kernels. Among the options, the RBF kernel is considered a suitable first choice, when compared to the polynomial and sigmoid kernels [73]. Additionally, RBF kernels have fewer hyperparameters, which makes the tuning process faster. For these reasons, it was decided to exclude the polynomial and sigmoid kernels from further analysis.

3

State-of-the-Art

3.1 Affective Gaming

The majority of videgames aim to provide entertaining and immersive experiences to players. However, designing and implementing a videogame capable of keeping the player engaged and attentive for long periods of time is often met with barriers associated with player generalisation, gameplay repetitiveness and inadequate progression [74]. Adaptive player-centric gameplay attempts to overcome some of these limitations, by establishing the player as the centerpiece of the game and generating adaptation strategies based on gameplay [21, 22]. These models evaluate a set of game metrics as input, for instance, *time to complete a level* or *the number of player deaths*, and use that information to adapt, for example the difficulty of the game, in an attempt to match the skill level of the player and maintain motivation. Although player-centric models introduce evident advantages compared to the static, non-adaptive gameplay, videogames are emotional experiences and can elicit a wide range of emotions in players. This, in turn, determines how the player interacts and performs in-game, defining the overall gameplay experience [12]. Thus, performance alone is insufficient to characterise the player experience. In that sense, AC opened a window of opportunity to explore new ways of interaction between videogames and players, and enhance the gaming experience, by integrating emotion into the design and development of videogames [75].

As one of the research branches of affective computing, affective gaming has been previously described as *"games in which the players' behavior directly affects the game objectives and gameplay"* [76].Thus, affective games differ from other videogames by the ability to adapt based on the affective states of the players, i.e., the affective state of the player causes some changes in the game, which in turn affect the players emotions [77]. As for how the game detects the player's affective state, the authors of [77] propose a taxonomy that organizes the type of feedback used in affective games into either direct or indirect. They describe direct feedback as the analysis of physiological reactions collected through biosensors applied to the player. A vast number of successful attempts have been made

in the field of physiology-based affective gaming. For instance, Yang et al. [78] have proposed that emotional states represented in the valence-arousal space, can be modelled through ECG, EDA, Electromyography (EMG) and skin temperature sensors, reporting an f1-score of 60.2% for valence and 57.3% for arousal detection, using a Linear SVM. Moreover, Chanel et. al. [14] carried out a study where they used Electroencephalography (EEG), GSR, respiration, skin temperature and Blood Volume Pulse (BVP) to detect three discrete emotional states (boredom, engagement and anxiety). They reported an overall accuracy of 63% obtained through the fusion of all modalities. Nogueira et. al [79] proposed a survival horror game to elicit changes in the valence and arousal of the players, which were detected with a rate of 78% and 85% for valence and arousal, respectively, using EDA, facial EMG and BVP. The methodology used in [80] resorts to a car racing game to evaluate player engagement at three different levels (low, medium and high), by measuring ECG, EDA, respiration, EMG and temperature. A recognition rate of 66% is reported using Linear Discriminant Analysis (LDA). Although, direct feedback metrics seem to be the most common choice for emotion recognition in affective games, possibly due to their nature that allows the continuous collection of unbiased data, their biggest disadvantage is the need to apply sensors to the player to monitor their physiological activity, which can generate discomfort or distract them from the game, and thus influence their affective state [77].

Parallel to direct feedback, indirect feedback attempts to overcome the previous shortcoming, by inferring the affective state of the player without using physiological data. The most direct and simple method is to ask the players themselves and build adaptation accordingly. Although it can produce very accurate player models, the intrusiveness of the method, together with the existence of experimental noise due to self-misjudgement (i.e., human factor), make data analysis difficult [81]. An alternative method is to analyse the player's interaction with the game's hardware and mechanics. As a way of analyzing the emotional states of the players, this methodology uses, for example, information related to the pressure applied to the buttons of the input device, the movement of the game avatar and their respective actions [77]. Frommel et al. [82] investigated emotion recognition through the combination of in-game performance and information gathered from the player interaction with the input device, and its applicability to a natural game context. The proposed game, named *Hiramon* consisted of writing Japanese hiragana characters on a pressure tablet. Several variables were evaluated from the interaction with the device (e.g., pressure, x and y position, duration of gesture) and used to compute statistical features. These, along with performance metrics, were then utilized to predict the intensity of valence, arousal and dominance. The study reported an f1-score of 57.7% for valence, 56.9% for arousal and 56.7% for dominance, using a Random Forest (RF) classifier. Despite being a more unobstructed method of assessing players' affective states, the downside of indirect feedback is that correctly defining and analyzing the gathered data can be time-consuming and not necessarily as precise as physiological signals [77].

The design of affective games has also been a target of some approaches. Gilleade et. at. [83] proposed a set of high-level heuristics for the design of affective games: *assist me*, *challenge me* and *emote me*. The first heuristic, "*assist me*", proposes a solution to players' frustration, for instance motivated by the inability to advance due to difficulty or missed clues, by measuring the physiological activity of the players and combining it with the game context to provide adequate adjustment to the game. The "*challenge me*"heuristic, on the other hand, tackles the limitation of most commercial games related to difficulty selection. Usually, players are presented with the possibility of selecting the desired difficulty between some predefined values (e.g., easy, medium, hard). In respect to the selected level, the difficulty then increases linearly or step-wise in the course of the game [84], attempting to match the player's skill level. This implies that features such as frequency of enemies, starting level or drop rates can only be set manually. Since the parameters associated to each of the levels are fixed, this method tries to map the same learning curve for all players. This generalisation can lessen fun and result in a negative experience. The solution proposed is to dynamically alter the level of challenge based on the players emotional response. Finally, the "*emote me*"heuristic consists of measuring the emotional state of the player and changing the game content to provoke the desired emotions as intended by the game designers. It is important to note that the design of affective games does not require the simultaneous implementation of these three heuristics. In fact, each one can be used in order to incite and emotionally motivate players to interact with the game and each can be explored and used according to the game's typology. In particular, the work of this dissertation will focus on the "*challenge me*"heuristic, more specifically, it seeks to explore how the difficulty of a game can affect players emotionally. As discussed in the chapter on affective computing, for it to be possible that there is some kind of affective adaptation, it is essential to build computational models capable of identifying and responding appropriately to players' emotions. Thus, by exploring the relationship between difficulty and emotion, the current work intends to propose an artificial intelligence model that predicts difficulty according to the set of emotions elicited in the player. The next section of the document will introduce the concept of flow and how it can be applied to videogames to explore the relationship between players' emotions and game challenge.

3.2 Flow and Game Experience

3.2.1 Theory of Flow

As a consequence of the highly interactive nature of videogames, these can elicit a wide range of emotional states, however knowing all of them is not necessary to evaluate player's experience. Thus, in the field of affective gaming, game experience is often discussed in relation to the theory of flow. The concept of flow was introduced by Mihaly Csikszentmihalyi, who describes it as sense of full and motivated concentration in a task,

while making full use of the skills required to do it and with a great level of fun, satisfaction and productivity [13]. A key argument of Csikszentmihalyi's theory is that players experience flow when subjected to an adequate level of challenge, however the presence of challenge does not necessarily lead to flow. Therefore, if the challenge required by a given game task exceeds the skills of the player, the task becomes to overwhelming and it can cause anxiety and frustration. Alternatively, not enough challenge can induce boredom. Hence, for a game to ensure the maintenance of a flow state, it must offer the right balance between challenge and skill, otherwise it will lead to one of the above states. Both of which would hinder the player's ability to achieve a "flow experience", leading to less involvement, engagement and possibly interruption of the game [14]. The concept of flow is illustrated in figure 3.1, with the skill level of a person along the vertical axis and the difficulty of a task along the horizontal axis.

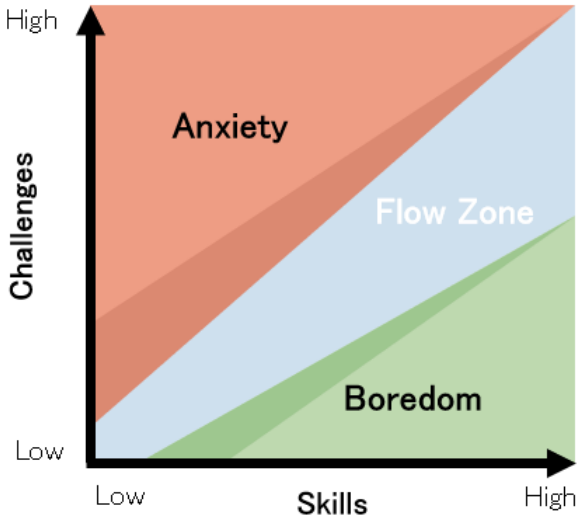

Figure 3.1: Flow diagram. Skill vs Challenge showing the "flow region"and regions associated with boredom and anxiety. The darker regions elaborate on the fuzzy boundaries of the "flow zone", which varies amongst players [85].

Generally, balancing these two components in videogames is challenging, since different players can have distinct approaches to the game, meaning players may experience and display different susceptibilities to either anxiety or boredom. For instance, some players may enjoy when difficulty slightly exceeds their skill level, others seek to be in control of the game and want challenge to be marginally under their abilities. This implies that the sense of flow is not identical to all players and has to be adjusted to the players goals, attitude and expertise [85].

In order for a game to be engaging and entertaining the level of challenge should match the player's abilities. This balancing has to be continuous, as the player's skill

level naturally increases as they continue playing. Based on this argument, the present work developed an approach similar to that of [14], where three emotional states of interest were defined, each one corresponding to a different region of the valence and arousal space, as presented in figure 3.2: negative valence and low arousal (boredom), positive valence and high arousal (engagement) and negative valence and high arousal (frustration).

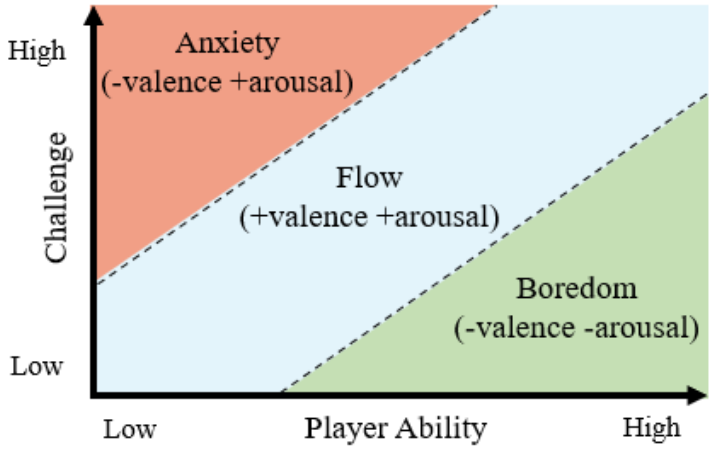

Figure 3.2: Flow diagram which was modified to cover three separate regions of the valence and arousal space. Adapted from [14].

Therefore, this work seeks to verify the validity and usefulness of the three states of emotion in predicting the level of difficulty experienced by players. For this purpose, a game was designed and implemented, in which the challenge is modulated by changing the difficulty level according to three levels (more on this in chapter 4). Nevertheless, it is relevant to mention that the "*challenge me*"heuristic presented in the previous section, is essentially the application of the theory of flow with affective-based adaptation of difficulty, commonly referred as Dynamic Difficulty Adjustment (DDA), in the field of affective gaming. Although the present work does not explore the concept of DDA in-depth, one of the long-term goals of the project is to use the proposed model to create a DDA mechanism, capable of providing adequate balance between ability and challenge. This has already been achieved by other studies in the area, for instance, Liu et al. [58] used a Decision Tree (DT) algorithm for adapting difficulty of a Pong game. They used numerous physiological signals such as ECG, EDA, EMG, and skin temperature to evaluate the anxiety of the player. They found that implementation of DDA through the use of an affective loop based on the level of anxiety was more effective than the traditional performance-based system, in terms of an immersive and challenging experience. On a similar note, Chanel and Lopes [86] designed and implemented a DDA system, which they used to adapt the difficulty of a Tetris game based on the level of arousal, measured through EDA. The proposed affective model used an ensemble deep neural-network to

predict anxiety and boredom with an overall accuracy of 73.2%. Thus, the present work differs from the examples above, since, on the one hand, it does not extend on the concept of DDA and, on the other hand, it seeks to explore VR as an affective medium.

3.2.2 Game Experience Assessment Methods

Aside from the theory of flow, studies also use self-reports as a method for evaluating the game experience. Self-reports provide a popular and accessible way for the psychometric measurement of emotions, by using questionnaires or interviews that rely on the player's self-report of their affective states. One such questionnaire is the Self-Assessment Manikin (SAM), proposed by Bradley and Lang in 1994 [87]. The SAM questionnaire, as illustrated in figure 3.3, measures valence, arousal and dominance, using three sets of five anthropomorphic figures which span along a 9-point scale. The participant marks the illustration they find the most fitting to their current state in each of the three scales. Although it was not originally developed to measure game experience, it has successfully been used, in several cases [14, 78, 79, 82], to assess the general affective state of the player after a gaming session.

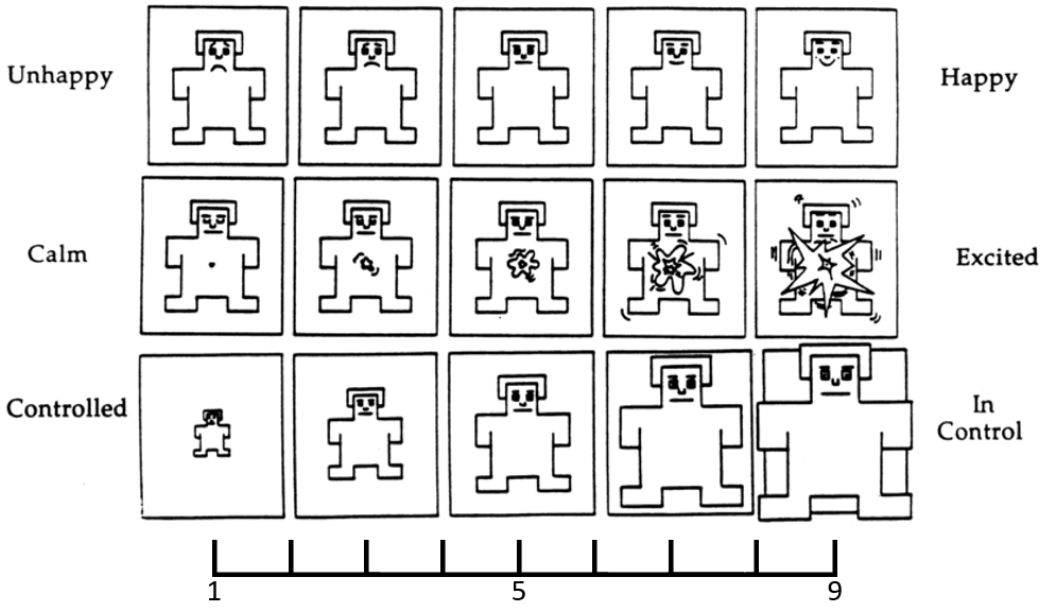

Figure 3.3: The Self-Assessment Manikin (SAM) Measure Scales [88]. Valence (first row), Arousal (second row) and Dominance (third row).

3.3 Virtual Reality

VR is a computer-aided design that, through synthetic sensory information, leads to the perception of simulated environments and their contents as if they were part of the real world [89]. Although, by nature, VR is "unreal", it manages to elicit emotions and

generate involvement through perceptual stimuli that include visual and auditory cues, but also, although less common, tactile and olfactory [90]. In the last two decades, there has been a significant increase in interest in using VR as a medium for videogames [91, 92], education [93, 94], rehabilitation and therapy [95–97], among others, as it offers a naturalness and connectivity to users, that other mediums lack. Additionally, VR also introduced itself into the field of emotion recognition, since, on one hand, it can be used as a more reliable elicitation agent to investigate human emotional behaviour in well-controlled designs [90], and one the other hand, it provides new ways of interaction in the context of HCI, where the assessment of emotions can enhance the communication with the end user application, for instance VR videogames [98].

Moreover, emotion elicitation protocols can benefit from VR in terms of immersion capacity [99], but also in other methodological aspects. Blascovich et al. [89] describes some of those aspects, two of which are transversal to the this study: ecological validity and lack of replication. The first aspect refers to the trade-off between experimental control and mundane realism, that is, the elicitation protocol must guarantee control over the experimental variables, and at the same time reflect similar situations to everyday life. This balance, while difficult to optimize, is a key factor in increasing participant involvement within well-controlled experimental settings [89]. Moreover, the use of VR allows to mitigate some of the limitations associated with the lack of replicability, for instance, researchers are often faced with the impossibility of perfectly replicating the experimental procedures of other studies, since both take place in physically different locations. Albeit the use of VR environments does not eliminate, in most cases, these methodological problems, it is expected to perform better comparatively to traditional non-immersive elicitation content [89].

Due to its ability to evoke target affective states, VR has been used as an affective medium for emotion elicitation in several studies [99]. Depending on the nature of the emotional-inducing material used in the study, the employed methodology can be categorized as passive or active [99]. Passive elicitation methods assess the subject as an observer of an emotional event. These mechanisms include watching images [100, 101], watching videos [102, 103], or being immersed in virtual environments [104–106]. In contrast, in active modalities the user is actively engaging with the emotional stimuli. These involve interactions with computer generated avatars, that promptly respond and adapt to the interaction through facial expressions and behavioural changes [107] or games that promote user engagement, entertainment and flow, through completing tasks and challenges [108, 109]. Broadly speaking, most of the literature related with emotion elicitation and recognition using VR, has focused on passive elicitation methods, which can be justified by the simplicity, cost-effectiveness, and flexible processing offered by this modality, along with the abundance of standards, norms and experimental practices readily available in the literature [99]. However, the use of audio-visual stimuli, without

any form of active interaction with the virtual environment, limits the full-dimensional experience and the impact it will have on the emotion eliciting process. Moreover, this methods are criticized by the lack of environmental validity, that potentially limits the generation of reliable theoretical frameworks [110]. Contrariwise to passive methods, active mechanism promote higher ecological validity and immersivity, which intensifies the emotional experience [99].

Furthermore, complex virtual environments, such as those associated with VR games, are believed to have the ability to induce a sense of presence in the virtual world, commonly referred as "*the sense of being there*"[111]. Under the effect of this feeling, users perceive themselves to be enveloped by the virtual environment to the point that virtual experiences can provoke similar perceptual reactions and emotions as those of the real world [111]. This ability to induce the feeling of presence in the user experience is seen as a major feature of VR. Although presence and immersion are sometimes used synonymously, when discussing their relation with videogames, the interchangeability between both concepts is only valid when the level of immersion, defined by Brown and Cairns [112] as the "the degree of involvement"on a videogame, is total. The authors describe total immersion as a state of complete detachment from reality, where players are entirely absorbed in the game, forgetting everything around them. It is therefore expected that, through the use of highly interactive synthetic environments, i.e., the combination of videogames and VR, it is possible to generate more intense emotional experiences. In fact, several studies have examined the interaction between presence and emotions in VR videogames, and all of them observed that a higher level immersion, i.e., presence, affects directly the vividness and intensity of the emotions experienced [93, 106, 113, 114]. Table 3.1, located at the end of this chapter, summarizes the main characteristics of works conducted in the literature using VR and Non-VR videogames for emotion elicitation and recognition.

3.4 Emotional Computerized Assessment

Machine Learning (ML) algorithms have been successively used in emotion related studies to discriminate between the user's affective states using several modalities, including physiological signals (as discussed in section 2.3). As described by Yannakakis et al. [12], the processing and interpretation of physiological data towards the detection and evaluation of the users emotional state, can be effectively achieved through the use of various supervised and unsupervised ML methods. Both categories base themselves in the construction of algorithms capable of mapping a set of physiological features to its' emotional state representation, by either classification or regression. The main difference between both comes from the fact that supervised algorithms use labeled data to train, i.e., the algorithm will evaluate a training dataset, where each element is labeled, and will establish a set of rules or mathematical models, on which it will base itself to evaluate and

make predictions using new data [115]. On the contrary, unsupervised learning models train on unlabeled data and have to recognize patterns in the data, in order to separate them, while simultaneously establishing the set of rules or mathematical equations necessary for the evaluation and prediction of new data, increasing the complexity of the models comparatively to supervised solutions [115]. Regarding the guidelines for the design and implementation of an affective system using biosignals, Munih et al. [116] presented a extensive overview of the general process of measuring, interpreting and using distinct ANS responses, necessary to create physiologically adaptive computing systems. As described in their work:

- The step 1) includes the selection of the appropriate psychological model. Such models refer to the discrete and continuous emotional models described in section 2.2, however, as highlighted by the authors, when system adaptation is desired, reworking of the models towards ad-hoc designs that include only the most relevant emotional states is usually preferable, since accounting for fewer states translates into the definition of fewer actions [116]

- The step 2) corresponds to the preparation of the training dataset, which is a requirement for supervised ML. In psychological computing, the training set refers to the set of psychological measurements (e.g., HRV, EDA) associated with induced psychological states (e.g., fear, anxiety, boredom), as such supervised methods learn by establishing associations between one and the other, which in turn requires knowledge of both the physiological measure and the respective psychological state [116]. Furthermore, the targeted emotional states must be present in the training dataset. As part of the adopted emotional model, these states need to be properly elicited in the subjects, as to ensure that the dataset contains useful information. The use of self-reports techniques (see section 3.2) is also recommended, to validate that the targeted emotions are successfully elicited, and ensuring a higher degree of reliability over the training data [116].

- In step 3), the measured raw physiological data of the training set is subject to extraction of features identified as relevant for the type of physiological measure [116]. In emotion recognition, feature extraction is understood as the transformation of the physiological signal into a set of attributes or characteristics, that are representative of the signal and suitable for a computational predictor [115]. Ciaccio [115] organizes features, often used in biomedical sciences, into two types: Physiological features and Statistical features. By definition, physiological features correspond to features that are derived from the available knowledge about the system under study, while statistical features are purely mathematical concepts that do not necessarily have a direct physiological meaning [115]. In terms of statistical features, these are usually computed under the assumption that physiological signals behave as independent Gaussian processes, and include mean and standard deviation as

descriptive statistics of the signal [115]. Although it is likely that the assumption of independency does not hold for physiological systems, since their behaviour is influenced by the same system (i.e., ANS), none of the studies reviewed for this dissertation tried to consider those dependencies during the feature extraction phase. On the other hand, several physiological features have been suggested for each of the biosignals used in this study. For instance, for ECG, HR and HRV are the most used time domain features. While in the frequency domain, Low Frequency (LF) band and in High Frequency (HF) band are commonly extracted [56]. Concerning EDA, some instances of the computed features are the number of SCR, average SCR and SCL amplitudes, maximum SCR and SCL amplitudes and SCR rise duration [14, 20, 51]. For respiration, the respiration rate is a commonly used time-domain feature [38, 117].

- The step 4), normalization or standardization, is not necessary but useful to reduce intra- and inter-subject variability [116].

- Next 5), in circumstances where there is an extensive number of features, all feature vectors are liable to a dimension reduction or feature selection. For instance, dimensionality reduction techniques generate a subspace of the original input space, creating new features from the existing ones and reducing the number of irrelevant and redundant dimensions, whereas feature selection selects the set of features that have a stronger relationship with the target variable, and discards features which have a weaker relationship or contribute little to the performance of the predictive model [118]. Both methods are employed to decrease cost and processing complexity [116, 118].

- Lastly 6), the data fusion stage will provide the classification or regression of the psychological state, using all available feature vectors originated from different physiological signals. On one hand, classification determines a discrete label for the feature vector (e.g., "boredom", "arousal"). On the other hand, regression methods assign a continuous value within a numerical interval to the psychological state.

When it comes to selecting the best algorithm, this is a process that depends on the objectives and dataset used. As stated by the No-Free-Lunch Theorem [119], there is no optimum machine learning method for either classification or regression tasks. In fact, in [116], the authors compared several supervised machine learning algorithms and concluded that, as accuracy and performance are dependent on the type and number of selected features, and vary in respect to context and the dataset, identification of the best algorithm is achieved by multiple testing and cross-validation on the same data pool.

Marín-Morales et al. [120] carried out an exploratory study in which they compared emotional responses during the exploration of the same space in a real and interactive

virtual environment. To this end, they used a SVM algorithm to classify valence and arousal, in terms of high or low, for each of the conditions, using EEG and ECG data. In the end, the model was able to predict valence, with an accuracy of 67% for the virtual environment, however arousal was removed due to low accuracy.

In their work, using ECG, EDA, EMG and RSP, Granato et al. [121] explored player involvement in a racing game, in which one version was in VR and the other was displayed on a monitor. They used several algorithms, SVM, Gradient Boosting (GB), Gaussian Process Regression (GPR), Naïve Bayes (NB) and K-Nearest Neighbors (KNN) to predict arousal and valence, and reported that SVM had the lowest Root Mean Square Error (RMSE) for both dimensions. Furthermore, they also evaluated differences in spatial presence for each of the versions, however it was inconclusive. Likewise, aiming to predict valence and arousal using EEG, Electrooculography (EOG), EMG, EDA, ECG, respiration, skin temperature, HR, BVP, Hinkle et al. [122] explored the use of various VR mini-games to stimulate different emotional responses. They evaluated the performance of SVM, NB and KNN algorithms and verified that the SVM presented the best performance without and with feature selection, 74% and 89%, respectively.

In another study, Moghimi et al. [123] investigated the usability and validity of using VR games to elicit different emotions, in terms of discrete (relaxed, content, happy, excited, angry, afraid, sad and bored) and continuous (positive valence - positive arousal, positive valence - highly positive arousal, negative valence - positive arousal and negative valence - negative arousal) models. By testing SVM, KNN, LDA and DT algorithms using features from EEG, GSR and HR data, it was found that KNN had the best overall performance in predicting both discrete (97%) and dimensional (96%) emotions. The SVM model came slightly under with a overall performance of 92% for discrete labels, and 92% for the valence-arousal clusters. However, none of the sessions managed to elicit sadness in the players, and this class was removed from the final evaluation.

Reidy et al. [94] propose an affective VR game for cognitive training, using facial EMG for emotion classification. Classification rates of 64% for valence and 76% for arousal were achieved through a combination of feature selection algorithms and a KNN classifier. From participants feedback, it was concluded that VR videogames offer a viable tool to promote engagement during cognitive training. Shumailov et al. [124] also use EMG to capture and recognize the affective states of players, during their interaction with a virtual environment. The study observed that EMG, placed on the lower arm, provides enough information to accurately measure valence (91%) and arousal (85%) using a SVM algorithm, with a RBF kernel. The purpose of the study carried out by Isahque et al. [125] was to analyze physiological responses (ECG, GSR and respiration), associated with the increase and decrease in stress, induced by a set of VR simulations and games. The study was able to accurately discriminate between stressful and relaxing stimulus using a GB

algorithm, with an accuracy of 85%, followed by a RBF-SVM with an accuracy of 76%. The study concluded that the VR fishing game, initially proposed to promote relaxation, was able to successfully reduce the stress induced in previous phases of the experimental protocol.

All reviewed studies (except the [43] study) performed user-independent emotion recognition. Although user dependent systems are, in most cases, easier to implement, and present better results, independent systems are considered more significant and applicable to the area of HCI [126]. Nonetheless, the following table compiles a non exhaustive review of the relevant studies concerning emotion elicitation using both VR and non-VR videogames, and their assessment using supervised ML models and physiological signals.

Table 3.1: A (non exhaustive) list of relevant studies concerning the use of VR and non-VR games to investigate emotion elicitation and assessment. The works were reviewed in terms of the algorithms used, material, biosignals, emotions models, best accuracy, system and number of participants. Note that best accuracy value is relative to the algorithm that is listed first in each row. (QDA: Quadratic Discriminant Analysis, LDA: Linear Discriminant Analysis, RSVM: RBF Support Vector Machine, KNN: K-Nearest Neighbors, LSVM: Linear Support Vector Machine, DT: Decision Tree, RF: Random Forest, RT: Random Tree, BNT: Bayesian Network, GB: Gradient Boosting, GPR: Gaussian Process Regression, NB: Naïve Bayes, EOG: Electrooculography, PPG: Photoplethysmogram, TEMP: skin temperature, RSP: Respiration).

Algorithm	**Materials**	**Signals**	**Emotion Classes**	**Best Accuracy**	**System**	**Participants**	**Study**
QDA, LDA, RSVM	Detect three discrete emotional states during a game of tetris with various levels of difficulty	EEG, TEMP, GSR, HR, BVP	Boredom, engagement and anxiety	63% (overall)	Non VR	20	[14]
LDA, KNN	Estimate player enjoyment using a car racing game	ECG, EDA, RSP, EMG and TEMP	Engagement (low, medium and high)	66% (overall)	Non VR	75	[80]
LSVM, RSVM, DT, RF	Investigate emotional responses during gameplay of a soccer game	ECG, EDA, EMG, TEMP	Valence and arousal	60% (valence), 57% (arousal)	Non VR	58	[78]
RT	Detect emotional responses during an horror game	SC, EMG and BVP	Valence and arousal	78% (valence), 85% (arousal)	Non VR	22	[79]
Ensemble XGBoost	Detect naturalistic expression of emotions using a mobile survival game	ECG, EMG, EDA, BVP, RSP	Valence and arousal	67% (valence), 69% (arousal)	Non VR	12	[43]
SVM	Classification of emotional arousal using video, music and games.	GSR, ECG, EOG, EEG, PPG	Excited and relaxed	89% (overall)	Non VR	5	[127]
SVM, KNN, RT, BNT	Classification of varying intensities (high, medium, low) of five affective states using a pong game	ECG, EMG, EDA, EMG, TEMP	Anxiety, boredom, engagement, frustration, anger	89% (anxiety), 84% (boredom), 84% (engagement), 83% (frustration), 89% (anger)	Non VR	15	[9]

SVM	Comparison of the emotional experience motivated by the exploration of a real vs virtual museum	EEG, ECG	Valence and arousal	67% (valence), N/A (arousal)	VR	60	[120]
LSVM, RSVM, RF, GB, GPR	Classification of players emotions in a VR racing game	ECG, EDA, EMG, RSP	Valence and arousal	N/A	VR	33	[121]
SVM, NB, KNN	Evaluation of the emotional response using multiple VR mini-games	EEG, EOG, EMG, EDA, ECG, RSP, TEMP, HR, BVP	Valence and arousal	89% (overall)	VR	5	[122]
KNN, SVM, LDA, DT	Classification of four valence and arousal clusters, and 8 emotional labels by using a boat simulator in VR	EEG, GSR, HR	Valence and arousal || relaxed, content, happy, excited, angry, afraid, sad, bored	95% (PVLA), 95% (PVHPA), 95% (NVPA), 91% (NVNA) || 96% (relaxed), 95% (content), 94% (happy), 96% (excited), 94% (angry), 93% (afraid), N/A (sad), 95% (bored)	VR	30	[123]
KNN, SVM, LDA	Estimating valence and arousal using a supermarkert VR game with different levels of difficulty	EMG	Valence and arousal	64.1% (valence), 76.2% (arousal)	VR	18	[94]
GB, LDA, DT, RSVM, NB	Classification of stress using a fishing game, a roller coaster simulation and a cognitive test	ECG, GSR, RSP	Stress and relaxation	85% (overall)	VR	14	[125]
SVM	Four VR mini-games were used to evaluate valence and arousal	EMG	Valence and arousal	85% (valence), 91% (arousal)	VR	8	[124]

4

Proposed System and Methodology

The experimental methodology adopted for this work is similar to that proposed and used in studies with respect to elicitation and recognition of emotions using videogames (see chapter 3) and can be summarized in the following steps: emotion elicitation; acquisition of physiological signals; physiological signals pre-processing; feature extraction; model selection; and classification.

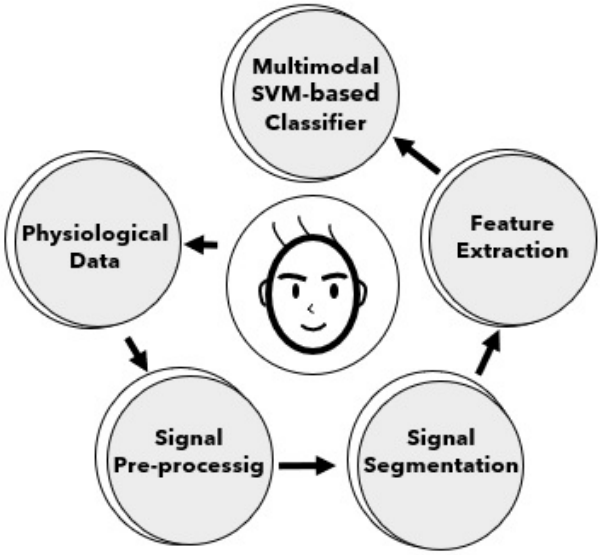

Figure 4.1: Proposed workflow for difficulty classification.

Figure 4.1 illustrates the proposed experimental procedure, which receives as input the physiological data, acquired from players during task completion, and outputs the level of difficulty, as predicted by the multimodal SVM-based classifier. Thus, this chapter is divided into 2 parts, the first of which will be dedicated to the description of the games' design and concept, followed by the second part which will further explore each of the steps illustrated above, by describing their practical implementation and underlining their respective importance for the general adopted workflow.

4.1 Game Proposal

4.1.1 The Wandering Druid

This work questions if it is possible to recognize player experienced difficulty through autonomous classification, while playing a series of challenging VR gaming scenarios. As such, a simple puzzle-based VR videogame was designed and implemented in Unity[1], and titled "*The Wandering Druid*". *The Wandering Druid* is based on the Trail Making Test (TMT) [128], which is a standardized neuropsychological test used for evaluation of cognitive and executive functions, and assessment of cognitive dysfunction [128]. The test requires the subject to connect a sequence of 25 points in ascending order (e.g., using a pen and paper; or a computer screen), in the shortest time possible. There are two parts to the test, however, the proposed game takes its inspiration from the second part, in which test participants are prompted to connect a series of numerical and alphabetical points in ascending order, while alternating between the two, as illustrated in figure 4.2. The option for the second part of the test is mainly due to the increased cognitive difficulty of the tasks [128], compared to the first part, which makes it more suitable for challenging players without any cognitive dysfunction. Thus, the choice to use TMT, as the basis of the games' core system, is mainly supported by the following summarized criteria:

- It is characterized by a set of rules that are easily understood by the participants and can be easily incorporated into a virtual scenario;
- In terms of mechanics and rules, these are accessible and do not require previous experience with Virtual Reality or videogames to be learned;
- It provides a way to evaluate the effects of both physical (i.e., speed, accuracy, and coordination) and mental (i.e., memory, planning and task switching) challenge, through careful handling of the game rules;
- There are an abundance of standards and experimental practices about the TMT, readily available in the literature.

Subsequent sections will elaborate on each of these points, as well as describe the design features of the proposed game in more detail.

4.1.2 Game Design

There are a multitude of definitions and theories available that address what videogames are, but a detailed analysis of them is beyond the scope of this project. However, videogames

[1] https://unity.com/

are, aside from a cultural form, an art form, a narrative form, an education tool, and more, first and foremost games. In the book "The Art of Game Design: A Book of Lenses", Jesse Schell [129] defines games as:

> *"A game is a problem solving activity, approached with a playful attitude."* [129].

Under this definition, the author presented the four pillars of game design - the Elemental Tetrad -, which evaluates multiple perspectives and definitions of game design and summarizes them into four basic elements: Mechanics, Story, Aesthetics and Technology [129]. These four elements are hold as equally important to a good game, supporting each other as they work to provide an adequate experience to the player. Foremost, mechanics encompass the goals and rules of a game. While the goals are roughly the same as those previously defined for the TMT, the rules that substantiate the game mechanics of the *The Wandering Druid* require further elaboration. As such, rules can be defined as the set of statements and directions that must be followed within a given game in order for it to be played correctly [129]. These are established by the game designer and are fixed throughout the game. Moreover, it is the interaction between the rules that create the formal and objective system underlying any game. In this sense, Figure 4.2 summarizes the ruleset that characterizes *The Wandering Druid* and is a direct result of the adaptation of the original TMT rules to a videogame scenario.

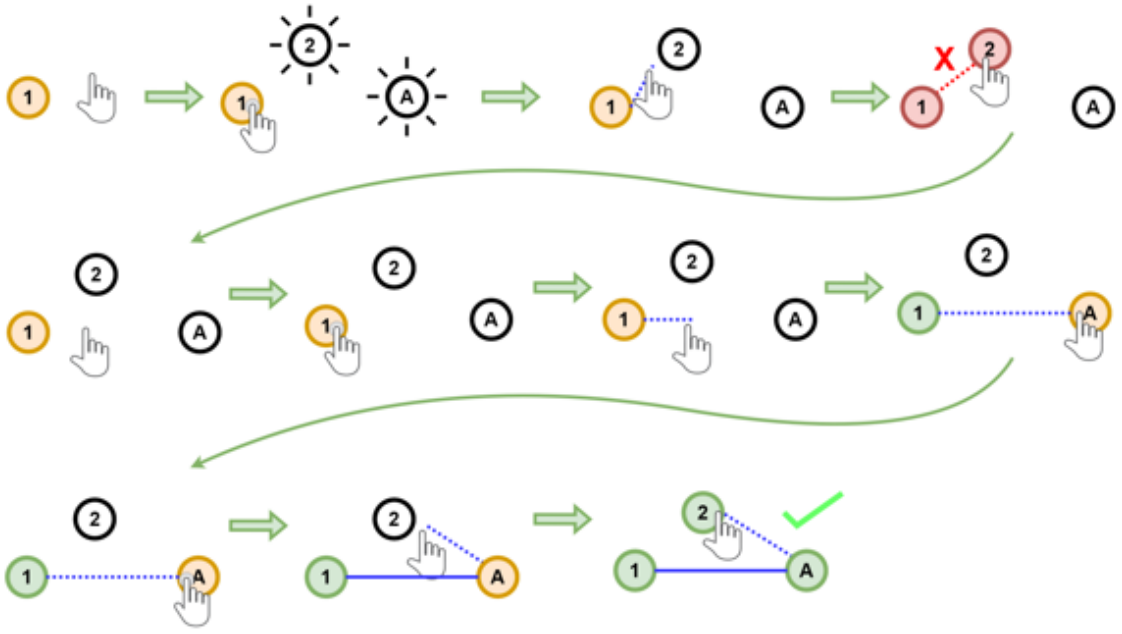

Figure 4.2: Summary representation of the *The Wandering Druid* VR game rules. Player must connect dots in an ascending order alternating between numerical and alphabetical symbols. Orange dots represents "the current active dot", or more precisely the next in-line to be connected. Red dots represents a wrong connection sequence. Green dots represents a successful connection.

Thus, the game itself presents the following core mechanics rules:

- Each sequence is completed by connecting a set of numerical and alphabetical points in ascending order, while switching between the two, for instance 1-A-2-B-3-C...
- At the start of each sequence, only point number 1 is displayed on the screen. The remaining points are hidden and only become visible once the player interacts with the former, after which they become permanently visible. This mechanic is repeated for all sequences.
- Players can only interact with the last successful point connected. If players fail a connection, they restart at the last successful point, maintaining their current progress.
- In case the player connects the dots correctly, the sequence is updated, otherwise the player is notified of the error and will have to try to connect to another point.
- The green points correspond to the points that have been successfully added to the sequence and can no longer be interacted with.
- Each sequence is characterized by a time limit. Players who don't complete the sequences under this rule will be moved to the next sequence. The timer is reset every new sequence and only becomes active after the first interaction of the player with the first point of the sequence (i.e., point 1).
- All sequence points are displayed in the same plane, since the games' implementation is intended to be as faithful as possible to the original core concept of TMT (i.e., there's no intention to explore the concept of depth within the game).

The players are introduced to the rules through a in-game tutorial. Below is a flowchart that summarizes the set of rules presented above, figure 4.3, and further describes how they interact with each other to define the flow of the game.

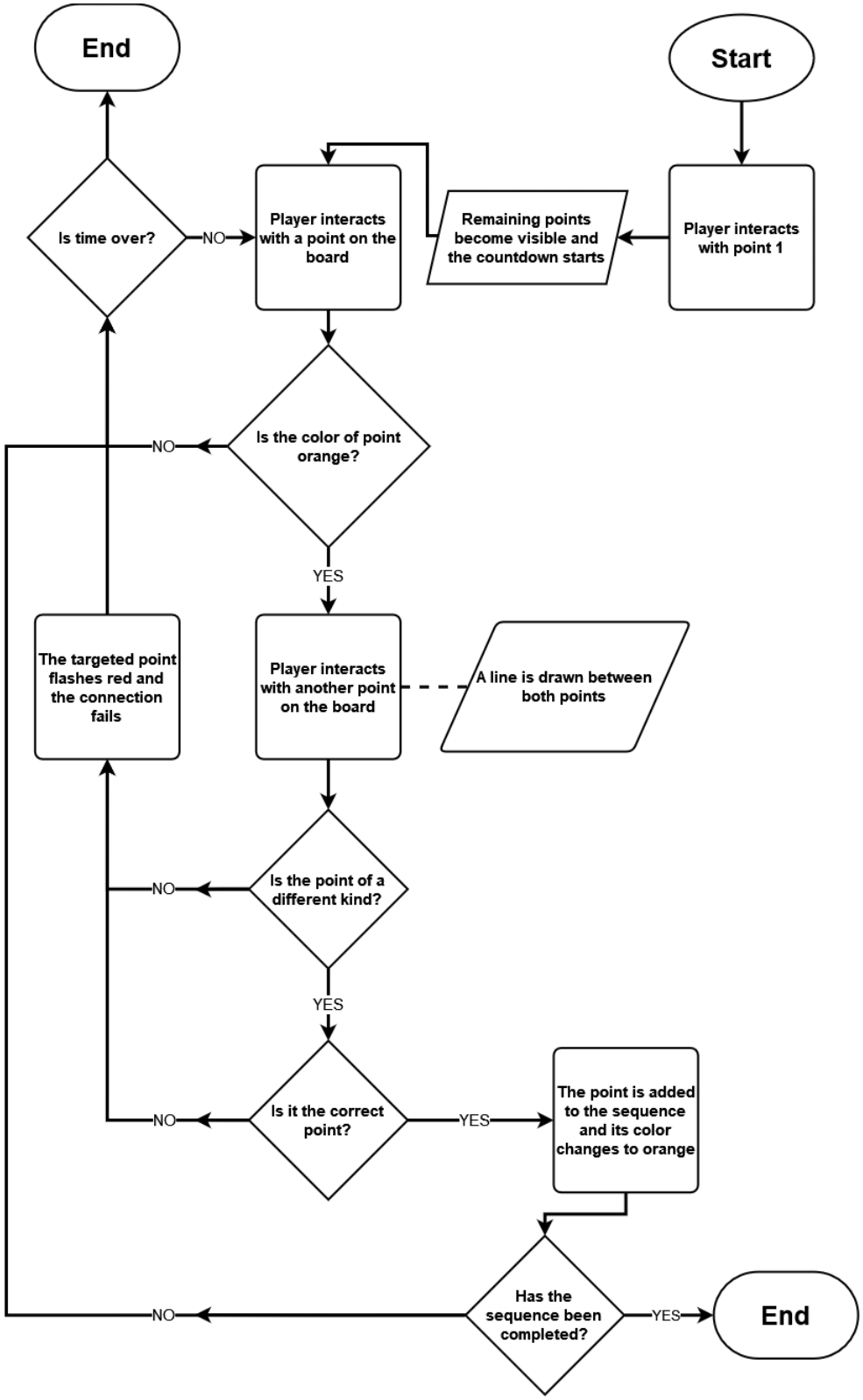

Figure 4.3: Flowchart describing the *The Wandering Druid* VR game flow.

In addition to the core systems design (i.e., definition of the basic rules of the game), it was necessary to construct a narrative (or story) around which to focus the game [129]. In *The Wandering Druid* the player assumes the role of a druid, during the Iron Age, who wanders from village to village aiding its citizens. In this version, players are tasked with the mission of summoning rain during a drought, by hand-drawing a set of line sequences, defined as magical runes, on a sheet of parchment, figure 4.4. In addition to the player, the village elder is also present, and whose role is to introduce the narrative premise to the player and present the game rules. For *The Wandering Druid* the designed player area corresponds to a 3D environment portraying an Iron Age hut. Sound wise, however, it was decided not to incorporate sound, as this would involve the use of headphones, and would make it difficult to communicate instructions from the experimenter to the participant during the game. Due to time constraints, and the complexity and scope of the project, narrative and aesthetics were designed with simplicity in mind, (i.e., as to reduce the amount of time required for implementation of the game), yet both are credible enough to convey to the player the idea that they are playing a videogame, and not just completing a cognitive test.

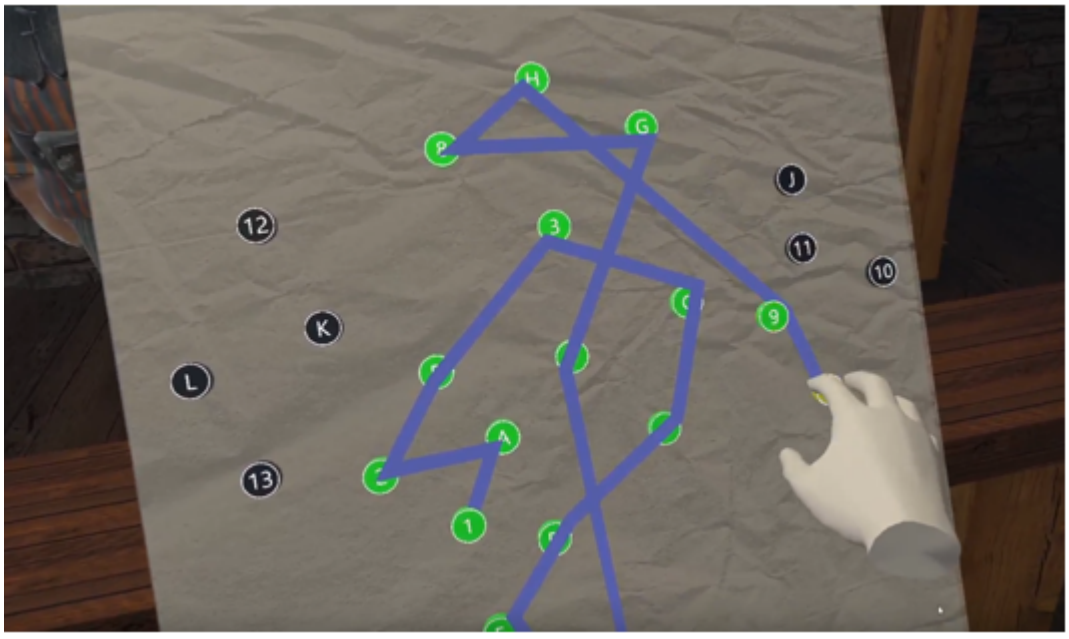

Figure 4.4: Screenshot of the sheet of parchment on which the player draws the sequences (i.e., the magical runes). Note that the interaction is limited to one plane, similar to the TMT.

Finally, the last element of the tetrad, technology, refers to what makes the game work, for instance the materials (i.e., hardware and software) and interactions with the other elements, therefore establishing what is, and isn't, possible [129]. Thus, the hardware chosen for *The Wandering Druid*, (i.e., aside from the basic setup necessary to run a game

like a computer or a console) was the HTC Vive Pro Eye[2]. This Head Mounted Display (HMD) fulfills the necessary criterion, in terms of player comfort, accessibility and privacy concerns. In terms of software, the game was implemented using the Unity Real-Time Development Platform, which provides complete solutions for both professionals and novices to create and operate real-time 3D experiences. This framework granted the versatility necessary to implement the desired features for *The Wandering Druid*; Moreover, the Unity ecosystem presents a wide range of high quality assets, preponderant for the adequate characterization of the play area.

To summarize, the Elemental Tetrad [129], used during the design process of *The Wandering Druid*, separates the basic elements of a videogame into four parts, making it easy to understand each element.

- **Aesthetics**: *The Wandering Druid*'s play area was designed to resemble an Iron Age Hut.
- **Mechanics**: The player must complete the sequences within a pre-established time limit, while complying with the rules of the game.
- **Story**: The narrative premise is that the player is a druid in a fantasy Iron Age village, tasked with the quest of summoning rain, by hand-drawing a set of magical runes.
- **Technology**: The game was implemented using the Unity Real-Time Development Platform and is played using the HTC Vive Pro Eye HMD.

4.1.3 Game Difficulty Parameterization

When discussing about the mechanical element of his model, Schell [129] also draws attention the strict relationship between player skill and in-game difficulty and how it should be adequately explored by designers to ensure players are in a state of flow. Since the objectives of this dissertation are strongly associated with the balance of both concepts, the following sections of the document will be dedicated to exploring its implementation in the context of the *The Wandering Druid*. Thus, to evaluate the effects of game difficulty over players emotions, three variables were initially proposed, i.e., *time limit*, *number of points displayed on the sheet of parchment* and *how often the lines intersect*, as a way to customize the difficulty of each sequence. Although the introduction of a time deficit in videogames has been proven as an efficient inducer of emotional stress [130], the specificity of *The Wandering Druid* required the proposition of additional variables for the parameterization of game difficulty. While time and number of points are easily understood, the last variable and its relation with the others is more dubious, hence requiring further clarification. As such, upon establishing the correct connection between

[2]https://www.vive.com/us/product/vive-pro-eye/overview/

two points, a line is drawn between them. This line can either interact with other lines or remain isolated. As all sequences are hand-drawn by the game designer, the manipulation of this variable, conjointly with the number of points, will allow for the definition of more and less strict heuristics when arranging the points on the sheet of parchment, hence creating levels and forms with different complexities.

It was initially hypothesised that by manipulating the three variables, it would be possible to shape the players' experience by exposing them to different levels of difficulty. On that account, the game was divided into three levels of difficulty - *Easy*, *Medium*, and *Hard* -, each comprised by a set of sequences that the player needed to complete in succession, and whose design was determined by the previous variables. As there was no prior knowledge of how long it would take each player to complete the sequences, it was necessary to carry out a pilot study to estimate the average time required to complete each level; Moreover, given the absence of a time limit and the heuristic nature of the difficulty parameterization, the study further explored the combined effect of the other two variables (i.e., number of points and the line intersection rate) over the players' perceived difficulty and in-game performance. Additionally, a fourth variable was also accounted for, in respect to the number of stages per level, which was set to a default value of five. Taking into consideration the last item, as the main purpose of the project strongly relies on the physiological data acquired during gameplay, each level should produce approximately the same amount of information, meaning players should spend roughly the same amount of time in each level.

4.1.4 Pilot Study

As mentioned earlier, a pilot study was carried out with the aim of answering the following research questions:

- How much time do players need, on average, to complete each difficulty level.
- Does the combined effect of both the number of points and the rate of intersection between lines, provide sufficient granularity to distinguish between the three levels of difficulty, based on players feedback and in-game performance.
- What is the appropriate number of stages (i.e., sequences) which need to be accounted for in each level, to ensure that all three levels generate roughly the same amount of physiological data.

4.1.4.1 Acquisition System and Procedures

Prior to the experiment, the participants had a adaptation period to the laboratory setting and were introduced to the experimental procedures. Their written consent to take part in the experience was also solicited, after which the experiment was initiated.

Inside the game, participants were first introduced to a tutorial, presented in text form, as a way to familiarize the players with the VR environment and the mechanics of the activities carried out during the experiment. Each participant then played the three difficulty variants, in a consecutive matter. For each of the levels, the total completion time was measured. This procedure was repeated for all three levels. The difficulty of the three levels was fixed for all participants and determined before the experiment, and each of the levels was designed according to the proposed heuristics described in table 4.1.

Table 4.1: Table of values initially proposed for the variables of time, number of points, intersection rate and number of stages for each level. The intersection rate corresponds to the ratio between the number of intersections and the total number of lines, for a given sequence.

Difficulty Variable	**Easy**	**Medium**	**Hard**
Time Limit (s)	Unlimited	Unlimited	Unlimited
Number of Points	13	16	25
Intersection Rate	0%	40%	40%
Number of Stages	5	5	5

The levels were preceded by an one minute break, in which participants were solicited to self annotate the level of difficulty experienced while performing the task, using a three items scale with the options *"Easy"*, *"Medium"* or *"Hard"*. Moreover, at the end of the rest period, participants were asked about how they felt and if they experienced any symptoms of cybersickness after playing. Such symptoms included vertigo, dizziness, nausea or headaches. Considering the participants' response, the experiment would either be resumed or terminated.

4.1.4.2 Participants

A total of 15 subjects, with ages between 20 and 35 years, participated in the pilot study, of whom 33% were female. This population was composed of elements that had no previous knowledge about the TMT, nor had any form of contact with the game before the experiment. All subjects participated as volunteers in the experiment and consented to the use of the collected data for the scientific purposes of this work. Regarding privacy, all the data was anonymized by assigning a numerical code to each user, and stored accordingly. No personal information was stored in electronic form.

4.1.4.3 Analysis and Results

The collected times totaled 45 samples, 15 for each level. In order to estimate the time that each participant spent, on average, in each stage, the recorded times were divided by the corresponding number of levels, as described in table 4.1. Additionally, to evaluate the performance of the participants for each level and to determine if the set of parameters chosen resulted in significant differences of performance between the three levels, it

was decided to calculate a performance metric corresponding to speed, by dividing the number of points of each level, as specified in table 4.1, by the average time of each participant, this way the time measures were converted to a common denominator.

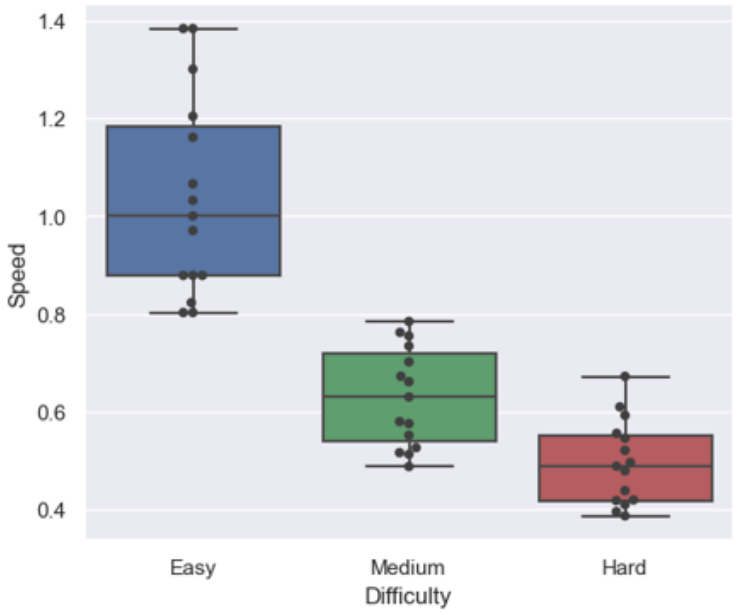

Figure 4.5: Boxplots of the participants speeds for each of the levels. The blue box corresponds to the speeds for the easy difficulty level ($MAD = 0.17$, $IQR = 0.3$), the green box is relative to the speeds for the medium difficulty level ($MAD = 0.09$, $IQR = 0.18$), and the red box contains the values of speed for the hard difficulty level ($MAD = 0.07$, $IQR = 0.13$), where MAD stands for Mean Absolute Deviation and IQR refers to the Interquartile Range.

The boxplots of the computed speeds for the three difficulty levels, as well as the corresponding empirical cumulative distribution functions (ECDF) were plotted (figures 4.5 and 4.6), and from the exploratory analysis of the data it was observed that:

- As difficulty increases, the speed at which participants complete the tasks decreases. This is to be expected since participants are faced with more cognitive demanding exercises at higher difficulty levels.

- There is a decrease in the variability of the performance metric, as difficulty increases. Note that, although difficulty for each task can be decomposed into 2 components: exterior (i.e., easiness of how participants interact with the VR environment) and interior (i.e., mostly related with the game itself, for instance, knowledge of rules, cognitive and motor acuity) - only the latter factors were considered when manipulating difficulty. To that extent, when evaluating performance for the easiest level, which should not be misclassified as control (i.e., absence of challenge), exterior factors are expected to be more prevalent, that is to say that, variability

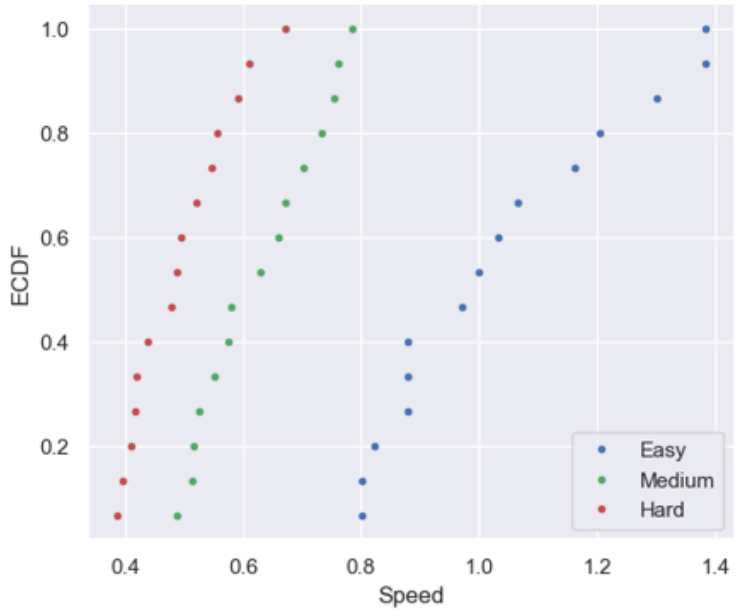

Figure 4.6: The empirical cumulative distribution functions for each of the difficulty levels.

amongst participants completion speed is mostly attributed to how comfortable they are with VR, since challenge, although present, is minimal. On the other hand, we argue that, as the difficulty increases, the weight that internal factors have on the difficulty outweighs the effects of the external factors. This means that, as tasks become more cognitive demanding, participants easiness with the VR environment will weigh less on performance. This, however, requires additional testing between participants with and without prior experience with VR.

- Finally, the boxplots corresponding to the medium and hard difficulties share a common subset of speed values, (i.e., their 95% confidence intervals intersect). This comes as no surprise, as most participants reported that the difficulty experienced between the medium and hard difficulties wasn't significantly different (75%). However, this hypothesis will be further evaluated using the performance metric (i.e., speed) previously computed from our dataset.

Table 4.2: Verification of the assumption of normality, a necessary condition for the application of parametric statistical tests. Asymmetry (skewness) and kurtosis were calculated. It was concluded that the normality assumption cannot be rejected.

Level	**Skewness**	**Kurtosis**
Easy	0.50	1.92
Medium	0.12	1.57
Hard	0.50	2.22

Previous to the application of any statistical test, it was necessary to evaluate the assumption of normality. Thus, for each level, kurtosis and skewness were computed. According to Hair et al. [131], data is deemed normal if the skewness is between −2 and +2 and the kurtosis is between −7 and +7. The results show that the data is right skewed, which is expected since for a variable such as speed, we expect to get a right skewed distribution, as there's no possibility that participants take less then zero seconds to complete the task. On the other hand, although kurtosis differs from the expected value of 3, for a normal distribution, it is still within the values defined in [131]. The results are presented in table table 4.2.

These conclusions are supported by the Q-Q plots of speed for each level (figure 4.7), which indicate a strong relationship between the observed value and the expected value, if the data was drawn from a normal distribution.

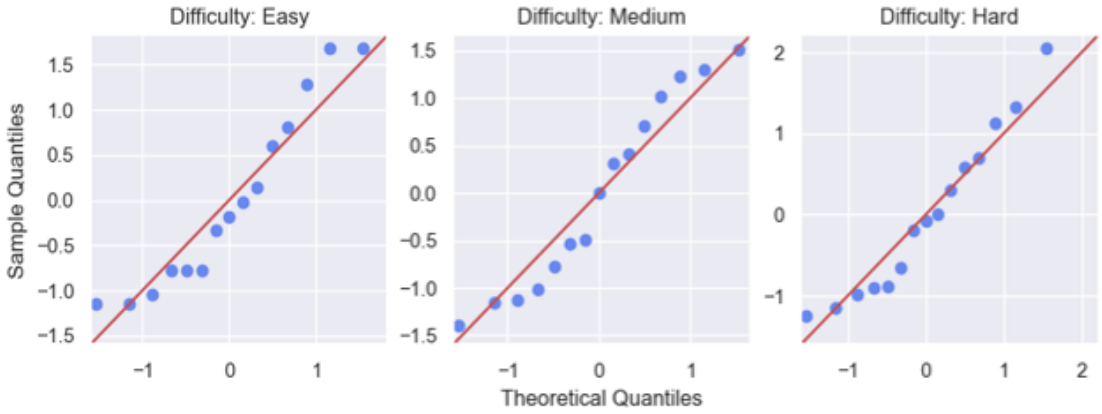

Figure 4.7: Normal Q-Q plots of the speeds of each difficulty level. The samples are sorted and plotted in relation to the quantiles calculated from a normal distribution. As can be seen, data is slightly right skewed, however the distribution of points relative to the normal distribution (red line) reveals a strong relationship between the observed values and the expected ones [132].

Under the assumption of normality, a alpha level was set to 5%, and a One-Way Repeated Measures ANOVA statistical test was applied to the speed data. A significant difference between the mean of, at least, two groups of speeds was observed ($F = 102$, $p << 0.05$). Furthermore, the assumption of sphericity was verified with a Mauchly's Test ($\chi^2 = 1$, $p = 0.15$) [133]. A post-hoc pairwise T-test was used to assess the existence of significant differences between pairs of groups and a Bonferroni correction [134] was applied to the results, to account for multiple comparisons. The results of the post-hoc test show that the mean speed for the easy level was significantly greater when compared to the mean speed of the medium ($T = 9$, $p_{corrected} << 0.016$) and hard ($T = 12$, $p_{corrected} << 0.016$) levels. This is observed from the boxplots, since the 95% confidence

interval for the easy level does not intersect with the intervals for the medium and hard levels. In addition, the test comparing the performance metrics for the hard and medium levels revealed the existence of a statistically significant difference between them ($T = 5$, $p_{corrected} = 0.000145$). In addition to evaluating the existence of differences in the difficulty of the various levels, regarding the chosen performance metric. The average time required to complete each level was calculated and the 95% confidence interval was estimated using bootstrap, table 4.3.

Table 4.3: Average completion times (by stage) for each of the difficulty levels, and the corresponding 95% confidence intervals, estimated using bootstrap.

Level	**Time (s)**	**95% Confidence Interval**	
		Lower Bound (s)	Upper Bound (s)
Easy	12.75	11.59	13.97
Medium	25.71	23.74	27.79
Hard	51.19	47	55.5

Taking into account the results of the pilot study, it is possible to answer the research questions proposed above. Thus, it was verified that the two variables proposed for the parameterization of the difficulty and their respective values, offer enough granularity to distinguish the different levels, in terms of performance. And, although participant reports do not point to a significant difference between difficulty, in particular between medium and hard difficulties, it is expected that the addition of a time limit variable will make these differences more evident. In addition, the values in table 4.1 have been updated, considering the values in table 4.3. Thus, it was decided to define an exaggerated time limit of 30 seconds for each sequence of the easy level, since it is intended that the addition of time does not affect the difficulty perceived by the majority of participants, on the other hand, a time limit of 30 seconds was defined for the medium level, since the objective is to challenge most participants, while ensuring that the percentage of completion is above 97.5%, and for the hard difficulty level, a time limit of 45 seconds was chosen, since the completion rate percentage is intended to be less than 2.5%. A updated version of table 4.1 is presented next.

Table 4.4: Table of the definitive values for the variables of time, number of points, intersection rate and number of stages for each level.

Difficulty Variable	**Easy**	**Medium**	**Hard**
Time Limit (s)	30	30	45
Number of Points	13	16	25
Intersection Rate	0%	40%	40%
Number of Stages	10	7	5

4.2 Experimental Settings

Subsequent sections will elaborate on the general steps presented at the beginning of the chapter, proposed for the elaboration of a system capable of predicting the difficulty experienced by players while playing *The Wandering Druid* VR game[3], through the measurement of their physiological responses.

4.2.1 Acquisition System and Experimental Setup

The acquisition system is composed by three modules: the sensor module used for the recording of the players physiological responses; the VR headset module; and the computer application module which ensures the communication between the various modules. The acquisition setup is illustrated in figure 4.8.

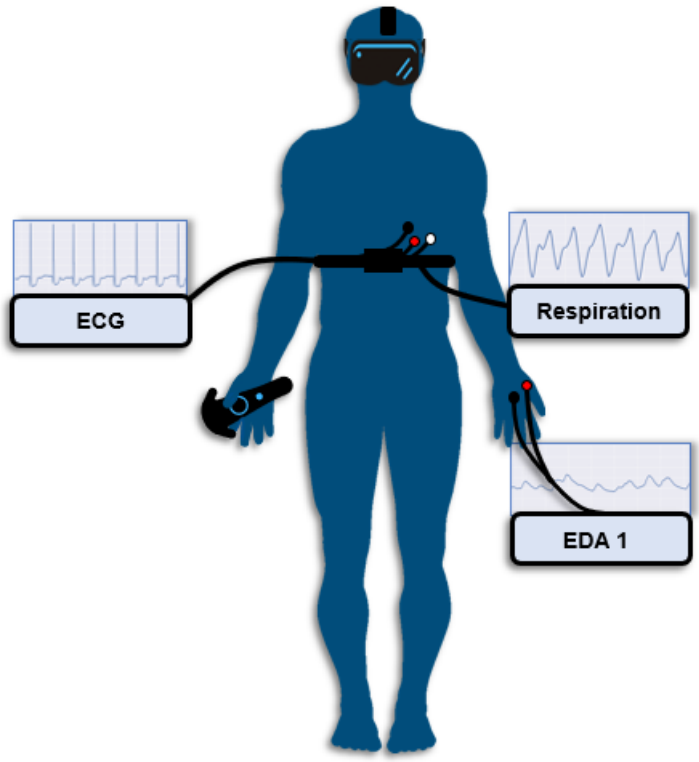

Figure 4.8: Illustration of the setup used during this experimental procedure.

A Biosignal Plux Explorer Kit[4], with a total of 4 channels, was used to collect the players physiological data during gameplay. More precisely, the EDA sensor was placed on the palm of the non-dominant hand, to ensure that the players dominant hand was left free to interact with the game and answer the questionnaires. In addition, the ECG sensor

[3] A video of the proposed game in action is available at: https://www.youtube.com/watch?v=zwi1RnEuCBc

[4] https://www.pluxbiosignals.com/products/copy-of-explorer

was placed on the chest, below the breast and slightly shifted to the left, relative to the external bone. As for the placement of the respiratory band, it was adapted according to the player's preference and the quality of the signal, having been placed either in the thoracic region or in the abdominal region.

Figure 4.9: HTC VIVE Pro Eye HMD and controller [135].

To finalize the experimental setup, the players used the HTC VIVE Pro Eye HMD (figure 4.9) and the associated motion controller (on the dominant hand), to interact with the system. Participants were asked to play in a seated position and to rest their non-dominant hand on their lap, in order to minimize the sources of motion artifacts. As previously mentioned, no audio stimuli were included in the final version of the game, to facilitate communication between the participant and the experimenter. Therefore, to minimize player distractions induced by noise from unforeseen sources, it was decided to carry out the experiment in a sound-insulated room.

4.2.2 Experimental Protocol

The experimental protocol used in this study is similar to that adopted for the pilot study (see section 4.1.4). As such, before the experiment began, participants had a period of adaptation to the laboratory environment, in which they were given a description of the study and its goals, along with a introduction to the experimental protocol. Any doubts or questions that the participant had were addressed during this period. At the start of the experiment, all participants had to sign a consent form, to ensure that they were aware of:

- the purpose and objectives of the study;
- the experimental procedure (i.e., duration of the experiment, description of the protocol, type of data recorded, and sensors used);
- the possible risks associated with the use o VR (i.e., cybersickness);

- the privacy issue (see next paragraph);
- the researchers contact in case of questions;

Regarding safety, participants were informed that the experiment did not present any risks and, if they wished to stop it, they could request it at any time during the experiment, ceasing it immediately. As for privacy, all the recordings were anonymized by assigning a numerical code to each participant, and stored accordingly (e.g., P1, P2, P3...).

After giving their consent, participants were solicited to fill-in a demographic questionnaire. Next, the wearable acquisition system was attached to the participants chest and hand, as described in the previous section. The HMD was placed on the participants head, who were instructed to remain seated for the duration of the experiment, and avoid moving the torso, and the non-dominant hand excessively, since this could impact the quality of the recorded signals.

The start of the game was preceded by a calibration period, in which two minutes of baseline activity were collected. During this period, the participant was asked to remain as static, and as quiet as possible. Once it finished, participants were introduced to the game tutorial. During this period, a Non-Playable Character (NPC) introduces the narrative premise, along with the rules and goals of the game. Additionally, before proceeding to the other phases of the game, participants have to complete a tutorial, to ensure they understand the mechanics of the game.

Once the tutorial was done, a total of three levels were played, each targeting a specific emotion, as shown in the table 4.5. As discussed in section 4.1.3, each of the levels presented players with different difficulties, and a distinct number of sequences to complete. Before starting any of the difficulty levels, players underwent a one minute rest period, the goal of which was to reduce the fatigue between each of the levels, and engross the participant into a neutral emotional state, minimizing the possibility of carrying the emotional state induced by the current level to the next.

To assess the success of the emotional elicitation process, participants were asked, after completing each of the levels, to self annotate their emotional state, using the SAM (see section 3.2). Furthermore, participants were also requested to self annotate the level of difficulty experienced (figure 4.10), and report on their current physical fatigue, using an adaptation of the Fatigue Assessment Questionnaire proposed by [136], figure 4.11. All questionnaires used in this study were translated to the Portuguese language, and incorporated into to the VR environment, to avoid the participant having to remove the HMD to answer them. This study was approved by the Ethics and Deontology Committee for Scientific Research of the School of Psychology and Life Sciences of the Lusófona University, where the experiment took place.

Figure 4.10: Difficulty questionnaire used in this study. It evaluates the difficulty experienced during gameplay into either "easy", "medium"or "hard"difficulties.

Figure 4.11: Questionnaire used to evaluate the physical fatigue experienced by players at the end of each level. It's an adaptation of the Fatigue Assessment Questionnaire proposed by [136], and evaluates fatigue in a 11 point scale which varies from "Total Fatigue", to "No Fatigue".

4.3 Dataset Validation - Electrodermal Activity

Previous to the application of any signal preprocessing techniques, for denoising and outlier removal, it is essential to visually inspect the data in order to assess its overall quality and ensure a higher level of reliability over the dataset. In this work, some of the EDA signals showed abnormal behavior that made its use impossible. Although the majority of the signals recorded for this study presented a dynamic typical of a EDA (see section 2.4.1), such as the one illustrated in figure 4.12. It was evident from the visual inspection of the signals, that not all EDA signals of participants were reliable.

For instance, the EDA signals of two participants showed signs of saturation, figure 4.13 and 4.14. This behaviour, which renders the signal unusable, cannot be fully explained, yet one can identify some potential causes. For instance, while an effort was

Table 4.5: Protocol summary, which includes the system preparation, game levels played and their respective targeted emotion.

	Target Emotion	**Duration(s)**	**Level Description**
Reading and signing the consent form		60	
Completion of the demographic form		60	
Placement of sensors and HMD		120	
Baseline Recording		120	
Completion of the game tutorial		300	
Resting phase	Neutral	60	The player must complete 10 sequences of 13 points and with no line crossings, each within a 30 seconds limit.
Easy Level	Negative Valence Low Arousal	300	
Difficulty annotation		10	
SAM annotation		20	
Fatigue annotation		10	
Resting phase	Neutral	60	The player must complete 7 sequences of 16 points and with line crossings, each within a 30 seconds limit.
Medium Level	Positive Valence High Arousal	210	
Difficulty annotation		10	
SAM annotation		20	
Fatigue annotation		10	
Resting phase	Neutral	60	The player must complete 5 sequences of 25 points and with line crossings, each within a 45 seconds limit.
Hard Level	Negative Valence High Arousal	225	
Difficulty annotation		10	
SAM annotation		20	
Fatigue annotation		10	

made to ensure that laboratory conditions did not change between sessions, which included maintaining the room temperature of the laboratory at $20°C$, it is likely that some of the participants suffered from very high sweating activity at the beginning of the experiment, due to the fact that these were conducted during the summer months. This hypothesis is supported by the high tonic activity exhibited by some of the participants at the beginning of the experiment, which decreased as they progressed through the activity.

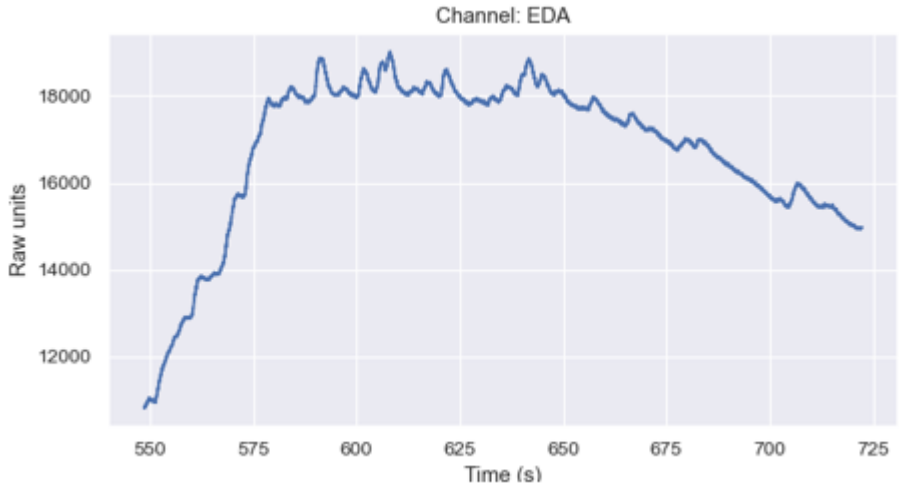

Figure 4.12: Expected behavior dynamics for an EDA (recorded on Participant 11, during gameplay of the Medium Difficulty Level).

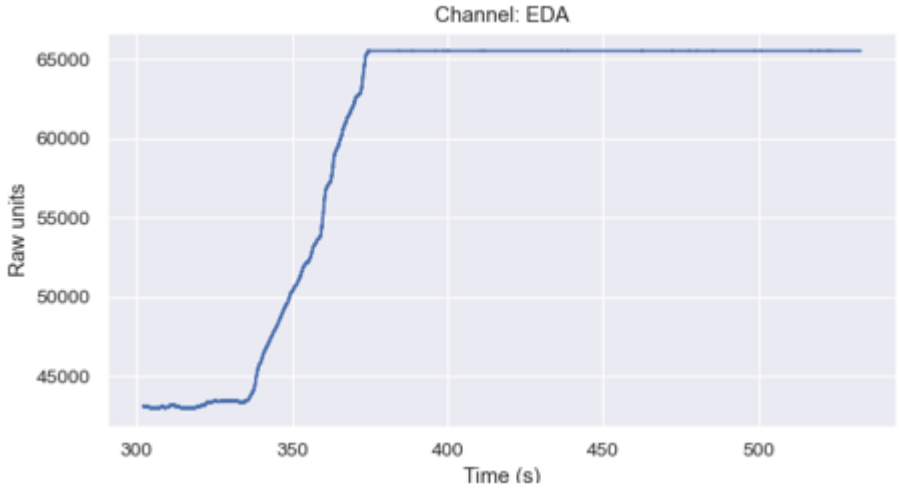

Figure 4.13: Abnormal EDA signal, due to saturation at the maximum ADC value (recorded on Participant 9, during gameplay of the Easy Difficulty Level)

In addition to saturation, excessive sweating can also be responsible for poor electrode contact. This was the case for three participants, whose EDA signals displayed abnormal curves, lacking any physiological meaning, figures 4.15 and 4.16.

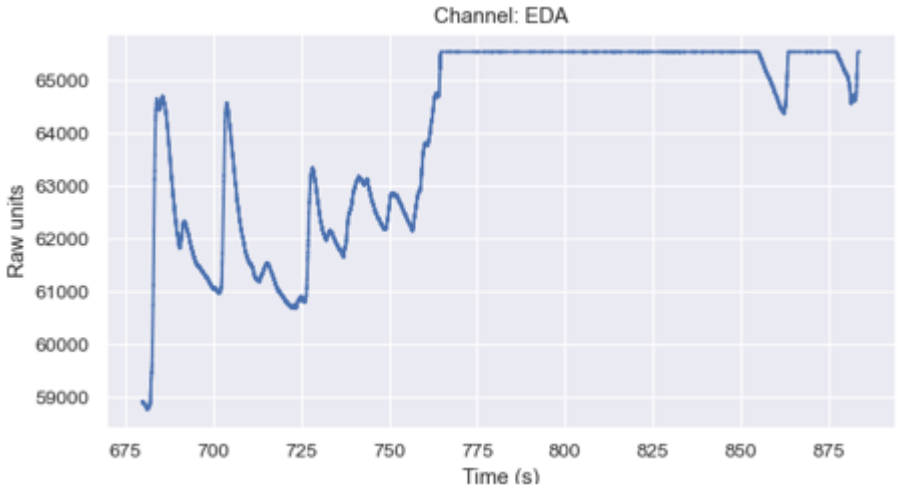

Figure 4.14: Abnormal EDA signal, due to saturation at the maximum ADC value (recorded on Participant 13, during gameplay of the Medium Difficulty Level)

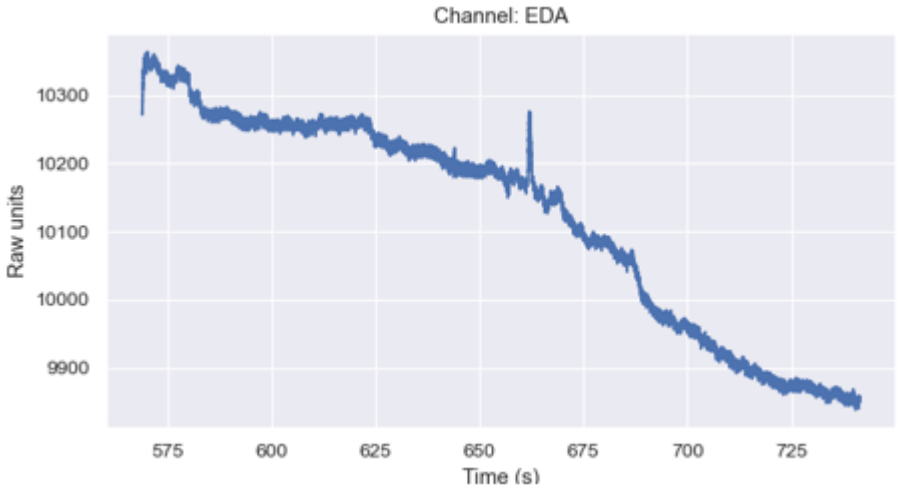

Figure 4.15: Abnormal EDA signal, due to poor electrode contact (recorded on Participant 18, during gameplay of the Medium Difficulty Level)

It was decided to consider, for the subsequent analysis, only signals that did not present the described anomalies, hence a total of five participants were removed from the initial dataset.

4.4 Signal Processing

In addition to the complex and subjective nature of physiological data, their high sensitivity to movement artifacts represents one of the greatest barriers to the development of affective systems for application in the real world. In fact, physiological signals are always contaminated by noise, whether it comes from electrostatic devices, muscle movements or other sources. This is particularly true in the case of the current study, where participants

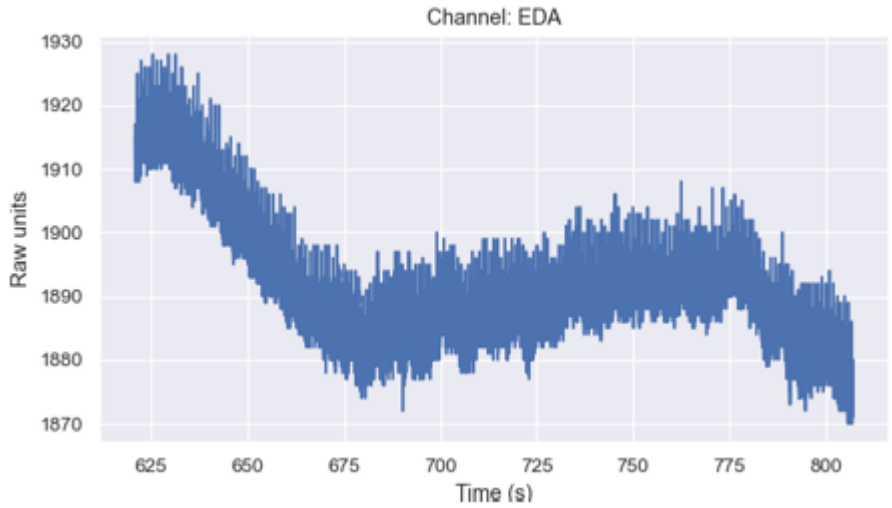

Figure 4.16: Abnormal EDA signal, due to poor electrode contact (recorded on Participant 28, during gameplay of the Medium Difficulty Level)

are tasked with playing a VR game, which requires them to move their bodies. This will inevitably lead to the addition of noise to the signals, and harden the task of unveiling the ground truth from the raw physiological data.

As a fundamental part of the workflow adopted for this work, the signal processing step consists of preparing the signals that will be used by the empirical models, both during training and prediction. Thus, this section will address the data conditioning techniques used for denoising and segmentation of the raw ECG, EDA and respiration collected during gameplay. Afterwards, the resulting signals move on to the next phase of the workflow, the feature extraction, as will be described in section 4.5.

4.4.1 Signal Filtering

4.4.1.1 Electrocardiogram Data

The ECG signal is considered a high-sensitivity physiological signal with a low frequency and low amplitude. In general terms, this signal is susceptible to corruption by various internal or external noise sources, which can include powerline interference, muscle movements, electrode–skin contact, motion artifacts, baseline wander, electronic and electromagnetic device interference, and respiration [56].

In this work, a kaiser window based Finite Impulse Response (FIR) filter was used to process the ECG signals [137]. Thus, a bandpass with a lower cutoff frequency of 3Hz, and a upper cutoff frequency of 45Hz was applied on the signals to suppress both low and high frequency noise. Figure 4.17 shows the filter response and an example of its application on a raw ECG signal.

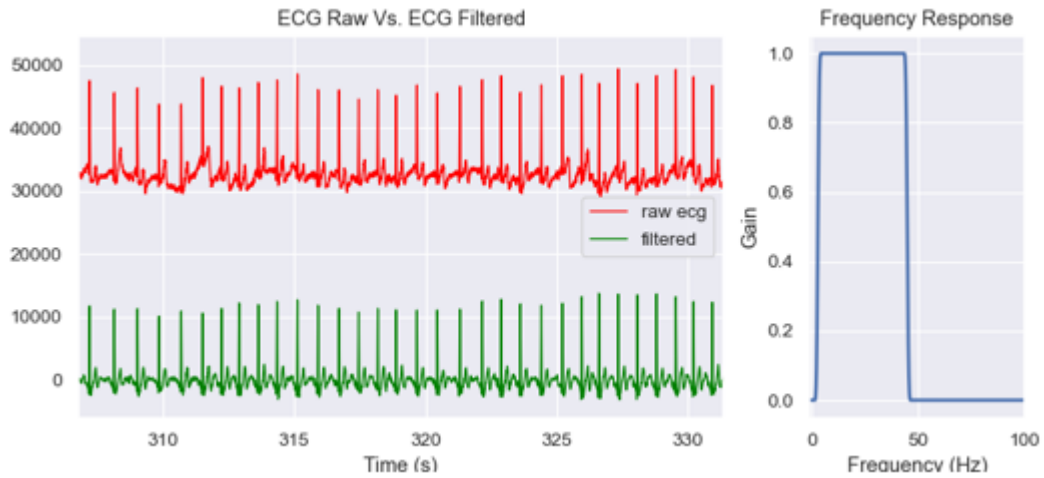

Figure 4.17: Frequency response of the proposed FIR filter and an example of its application to a raw ECG signal.

4.4.1.2 Electrodermal Activity Data

As described in section 2.4.1, the EDA signal is a relatively slow signal, which can be differentiated into a slowly varying tonic factor, and a fast varying phasic factor, ranging from 0.5Hz to 1.5Hz [138]. These signals are sensitive to various noise sources, which include physical movement, ambient temperature and humidity, and electrical noise [138, 139].

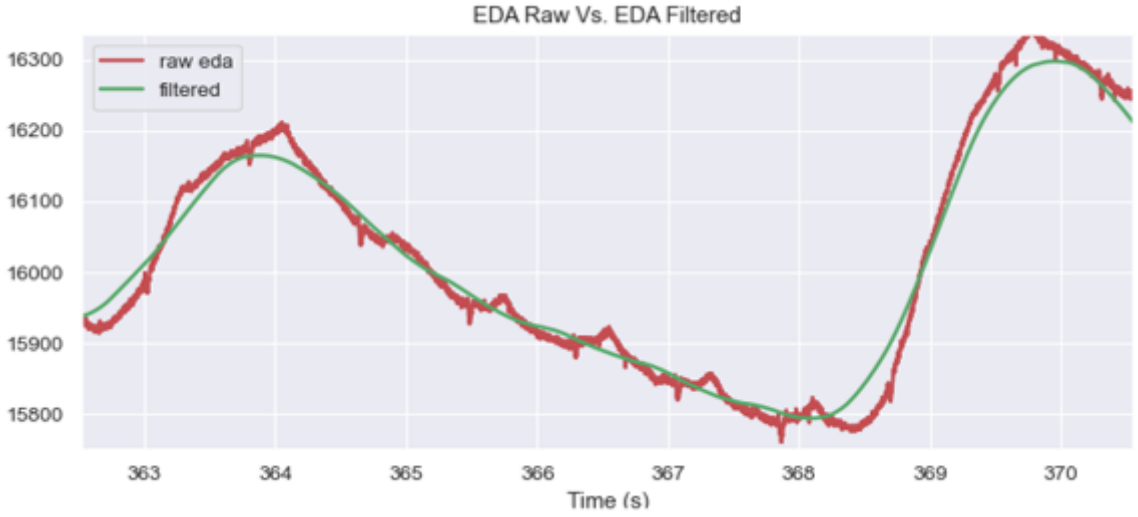

Figure 4.18: Example of a raw and filtered EDA signal.

The method adopted for filtering the EDA signals was based on the methodology proposed in [14]. Thus, a boxcar kernel of dimension 1024 (samples) was convoluted with the input signal, to smooth it, and eliminate the high frequency noise. Figure 4.18 shows an example of the filter in action.

4.4.1.3 Respiration Data

Due to the position of the band, and the nature of the tasks performed by the participants during the experiment, the respiration signal was affected by torso and arm movement artifacts.

To reduce the noise of the respiration signals, this work followed the approach used by [140]. Thus, a 4^{th} order low-pass butterworth filter with a cutoff frequency of 0.7Hz, was applied to the signal, followed by subtraction of the moving mean of the filtered signal, with a width of 4 seconds, to remove potential baseline wanders and large motion artifacts. An example of the application of the filter is shown in figure 4.19.

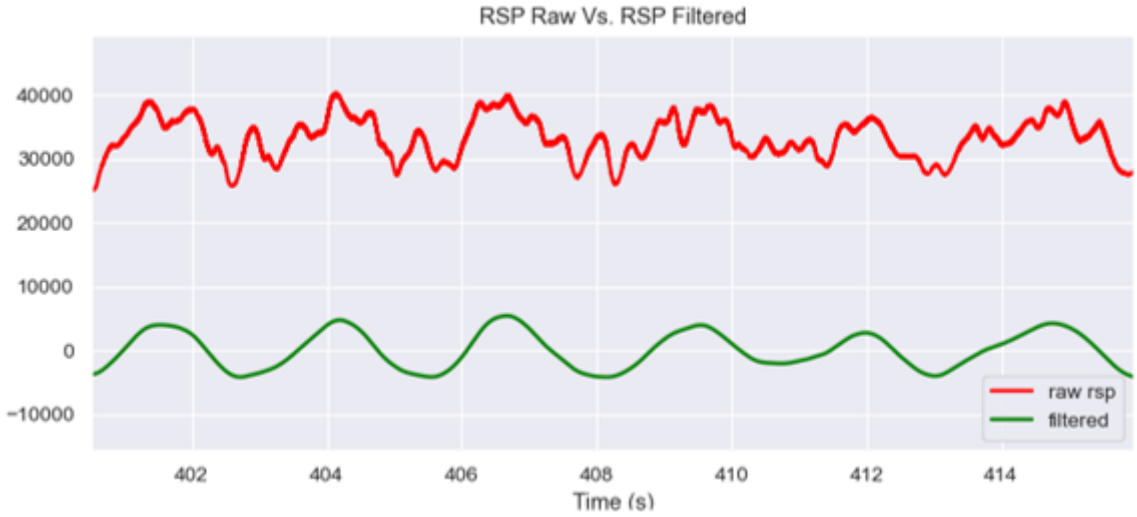

Figure 4.19: Example of a raw and filtered respiration signal.

4.4.2 Data Segmentation

The segmentation process of the three physiological signals is a fundamental step of the adopted processing pipeline, and corresponds to the identification of the templates on which the feature extraction techniques will be applied. In this work, segmentation of the signals was performed in two consecutive moments. The first of which required the definition of temporal markers that allowed the synchronization of game events with the acquired signals, and that facilitated their subsequent segmentation. These markers were built into the game and associated with specific game segments, e.g., the beginning and ending of each level. This saved the experimenter from the trouble of having to manually mark the portions of the signal that mattered, as this was done automatically by the game.

Afterwards, the portions of the segmented signal deemed relevant, were subjected to a second segmentation step, whose characterization is specific for each signal. In the case of the EDA, this was divided into the respective tonic and phasic components, and the

latter was used to identify the occurrence of SCRs. On the other hand, considering the periodicity of the ECG and respiration signals, segmentation was applied on the level of their cycle wave, by detecting specific peaks, as will be described in the subsequent sections.

The computation of this processing step was performed mainly using the *NeuroKit2* [141] library, which provides several functions for processing physiological signals, using the *Python* language. In the case of the ECG, it was decided to implement an additional step to remove outliers, based on the method proposed by [142].

4.4.2.1 Electrocardiogram Data

To segment the ECG signals, the default R peak detection method implemented by *NeuroKit2* was used. This algorithm returns the indices of each peak, thus allowing to discriminate the signal according to the templates of the ECG. The method detects QRS complexes according to a steepness criterion applied to the absolute gradient of the ECG signal. Subsequently, R peaks are determined as local maxima in the QRS complexes. A minimum delay of 300ms between R peaks is enforced, so that R peaks that follow other R peaks by less than the established threshold are discarded. This method produced reliable results regarding the detection of R peaks, as exemplified in figure 4.20, where a total of 11 templates were correctly identified.

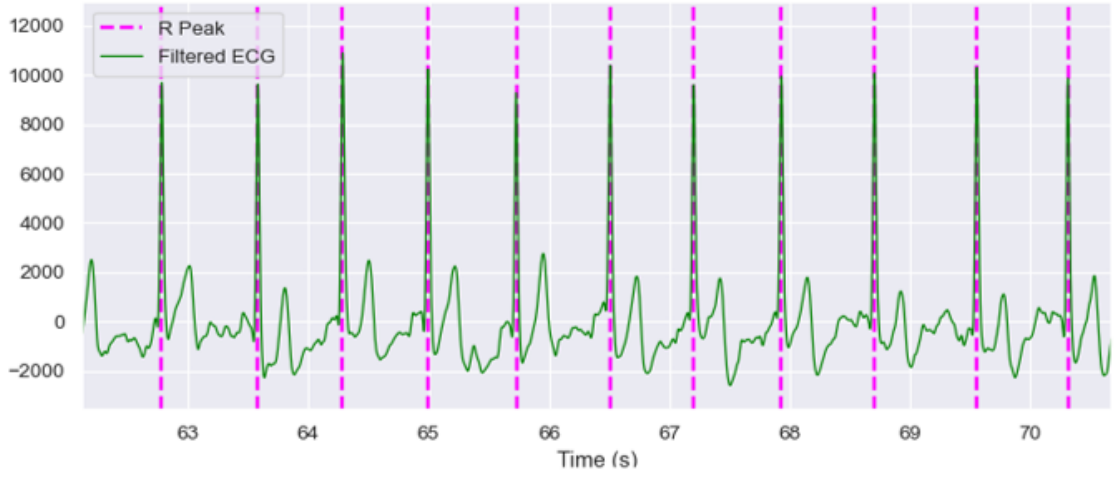

Figure 4.20: Identification of the R peaks, and discrimination of 11 templates, as a result of the segmentation step.

Although this method produced satisfactory results, it was necessary to apply a second algorithm to exclude abnormal templates identified by the method as R peaks. For this, the methodology proposed in [142], which considers the morphology of the QRS complex, was followed. It consisted of the estimation of the average template of the whole ECG,

determined through the average of all individual ECG cycles, segmented as blocks of [-200, +400]ms centered on the R peak, as illustrated in figure 4.21. Then, the distance of each ECG block, in relation to the average template is computed using the cosine distance as metric. Templates whose cosine distance is greater than the threshold established by the authors are discarded [142]. This method proved to be quite efficient in removing outliers.

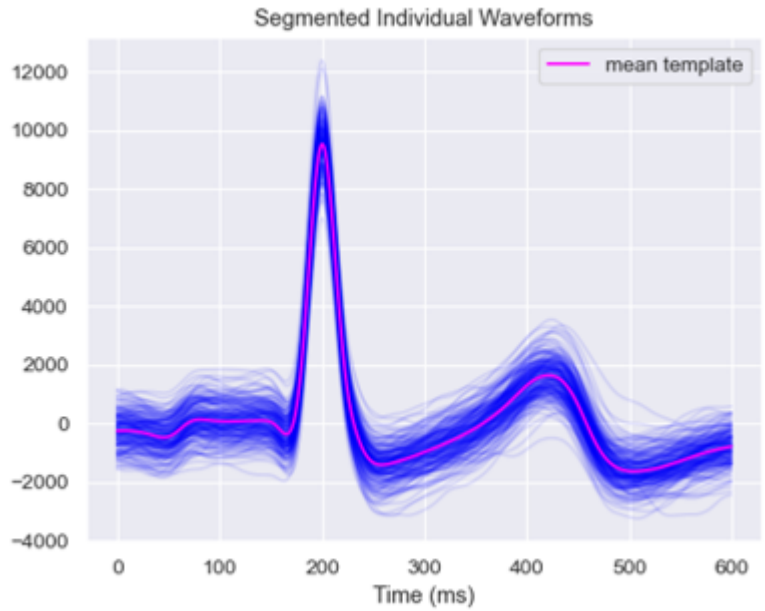

Figure 4.21: Segmented individual templates, used for the computation of the mean ECG template.

4.4.2.2 Electrodermal Activity Data

The segmentation of the EDA signal required its separation into the corresponding tonic and phasic components, in order to process the latter. For this, the *acqknowledge* method included in the *NeuroKit2* package was used. It consisted of computing the tonic component through the convolution of the EDA signal with a median kernel, which was then subtracted from the EDA signal to obtain the phasic component.

Processing of the phasic component and identification of SCR peaks was performed by detecting peaks whose relative height was higher than 10% of the maximum of the heights. The result of the segmentation pipeline adopted for the EDA signal can be seen in figure 4.22.

4.4.2.3 Respiration Data

Each respiration cycle starts with an inhalation period, that corresponds to an increase in the signal amplitude, and is followed by a exhalation period associated with a decrease

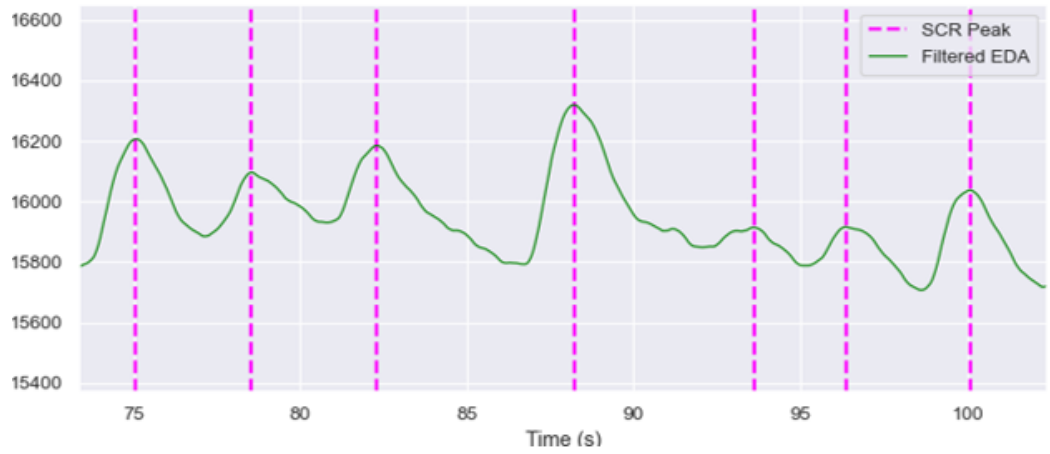

Figure 4.22: Identification of the SCR peaks, and discrimination of 7 templates, as a result of the segmentation step.

in amplitude. Thus, a cycle may be interpreted as the period between inhalations or troughs. Thus, to perform the segmentation of the respiration signal, this work followed the approach proposed by Khodadad et al. [143]. The method starts by detecting zero crossings in the raw signal. The result is then used to find extrema by searching minima (troughs) between falling zero crossing and rising zero crossing, and searching maxima (peaks) between rising zero crossing and falling zero crossing. Afterwards, the vertical distance of each extrema relative to its direct neighbor is calculated, and extrema whose absolute distance is less than the threshold determined in [143], are defined as outliers and excluded. Finally, an additional step is applied to make sure that the alternation of peaks and troughs remains consistent [143].

Figure 4.23 illustrates one respiration signal after this segmentation step, where 7 peaks were detected, each associated with the respective trough.

4.5 Feature Extraction and Windowing

4.5.1 Feature Extraction

Feature extraction is a key step for effective model construction, since it extracts meaning from the data. This section will be focused on the description of the analysis made regarding the choice of the appropriate features. These correspond to a set of inputs, representative of each physiological signal, on which the model will base itself to perform the classification tasks. Hence, to compute all the features from the ECG signal, several high-level functions implemented by *NeuroKit2* were used. For the extraction of features from the EDA and respiration signals, it was necessary to implement separate methods that allowed the extraction of features identified in the literature as meaningful, but not

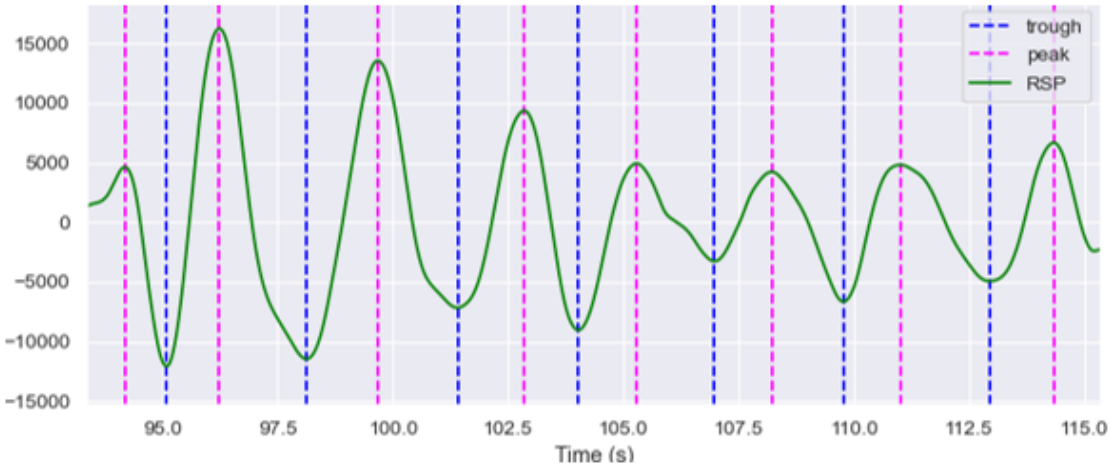

Figure 4.23: Identification of the peaks and troughs, and discrimination of 7 templates, as a result of the segmentation step.

available in the *NeuroKit2* package. Hence, a total of 63 features were computed from the three signals.

4.5.1.1 Electrocardiogram Data

A total of 18 features were extracted from the ECG signals using the R peaks identified during the segmentation step. These are categorized into time and frequency domain, and are regarded as potentially relevant indices of both sympathetic and parasympathetic activity [55]. Table 4.6 provides a summary description of each of the features.

Table 4.6: ECG features and their description [141].

Domain	Feature Name	Description
Time Domain	HR_Mean	Mean of the heart rate
	HR_SD	Standard deviation of heart rate intervals
	HRV_MeanNN	Mean of the RR intervals
	HRV_SDNN	Standard deviation of RR intervals
	HRV_RMSSD	Root mean square of successive RR interval differences
	HRV_SDSD	The standard deviation of the successive differences between RR intervals
	HRV_CVNN	The standard deviation of the RR intervals (SDNN) divided by the mean of the RR intervals (MeanNN)
	HRV_CVSD	The root mean square of the sum of successive differences (RMSSD) divided by the mean of the RR intervals (MeanNN)
	HRV_MedianNN	The median of the absolute values of the successive differences between RR intervals (ms)

	HRV_MadNN	The median absolute deviation of the RR intervals
	HRV_MCVNN	The median absolute deviation of the RR intervals (MadNN) divided by the median of the absolute differences of their successive differences (MedianNN)
	HRV_IQRNN	The interquartile range (IQR) of the RR intervals
	HRV_pNN50	Percentage of successive RR intervals that differ by more than 50ms
	HRV_pNN20	The percentage of RR intervals greater than 20ms, out of the total number of RR intervals
Frequency Domain	HRV_LF	The spectral power density pertaining to low frequency band (from 0.04 to 0.15Hz)
	HRV_HF	The spectral power density pertaining to high frequency band (from 0.15 to 0.4Hz)
	HRV_LFn	The normalized low frequency, obtained by dividing the low frequency power by the total power
	HRV_HFn	The normalized high frequency, obtained by dividing the high frequency power by the total power

4.5.1.2 Electrodermal Activity Data

Concerning the physiologically distinct attributes of the EDA signal, different features were computed for each of the components. As such, five statistical features were computed for both the tonic and phasic elements of the EDA, which include mean, standard deviation, median absolute deviation, median and area under the curve (AUC). Whereas, in the particular case of SCR, four additional features were extracted, regarding the number of SCR peaks, average rise and recovery times, and the maximum response amplitude. Table 4.7 summarizes the features used in this work, regarding the EDA signal.

Table 4.7: EDA features and their description.

Component	**Feature Name**	**Description**
Tonic	SCL_Mean	Mean amplitude of the tonic component
	SCL_SD	Standard deviation of the amplitude of the tonic component
	SCL_MAD	Median absolute deviation of the amplitude of the tonic component
	SCL_Median	Median of the amplitude of the tonic component
	SCL_AUC	Area under the tonic curve
Phasic	SCR_Mean	Mean amplitude of the phasic component
	SCR_SD	Standard deviation of the amplitude of the phasic component
	SCR_MAD	Median absolute deviation of the amplitude of the phasic component
	SCR_Median	Median of the amplitude of the phasic component
	SCR_AUC	Area under the phasic curve
	NP	Number of SCR peaks
	RET_Mean	Mean rise time of the SCRs (i.e., the time it takes for SCR to reach peak amplitude from onset)
	RIT_Mean	Mean recovery time of the SCRs (i.e., the time it takes for SCR to decrease to half amplitude)
	RA_Max	Maximum SCR amplitude

4.5.1.3 Respiration Data

For the respiration signal, with the exception of the respiration rate, four time-domain attributes were initially computed for each cycle wave. These were the inhalation duration, exhalation duration, respiration amplitude, ratio of inhalation/exhalation times and first order differences of the exhalation duration. For each of them, the mean, median, standard deviation, absolute median deviation and 80% percentile were calculated and used as features. Additionally, frequency-domain specific features were extracted, similarly to the ECG signal. Table 4.8 provides a description of the respiration features used.

Table 4.8: Respiration features and their description.

Domain	Feature Name	Description
Time Domain	RR	Respiration rate
	ID*	Inhalation duration corresponds to the time elapsed from a valley of a signal to the next peak, which denotes the maximum expansion of the chest in the respiration cycle
	ED*	Exhalation duration corresponds to the time duration between the peak and the next valley
	Range*	Difference between the amplitude of the peak and the minimum amplitude the signal attains within a respiration cycle
	IE*	Ratio of inhalation duration to the exhalation duration of a respiration cycle
	FDE*	First Difference of Exhalation is derived by computing the first order differences of the exhalation durations
Frequency Domain	RSP_LF	The spectral power density pertaining to low frequency band (from 0.04 to 0.15Hz)
	RSP_HF	The spectral power density pertaining to high frequency band (from 0.15 to 0.4Hz)
	RSP_LFn	The normalized low frequency, obtained by dividing the low frequency power by the total power
	RSP_HFn	The normalized high frequency, obtained by dividing the high frequency power by the total power

*The mean, median, standard deviation, absolute median deviation and the 80% percentile were calculated for this feature.

4.5.1.4 General Considerations

Much of the work done for this dissertation, and in particular for feature selection, is exploratory in the sense that, although a wide range of features have been proposed as valid indices for emotion recognition, there is no "best"set of features. Hence, statistical characteristics, such as those selected for this work, are considered standard, as they are frequently used in emotion recognition studies. In addition, more advanced features, such as those taken from the ECG frequency domain, are also commonly used. So the choice of features for this work was based on a criterion of popularity and computational simplicity.

4.5.2 Windowing

The extraction of features from the signals was performed according to two different procedures, the first one computes the features from the entire signal recorded for each level. Whereas, the second breaks the signal of each level into windows and from each one extracts the set of features described above. This technique, termed windowing, segments the signals according to a time window of fixed duration. It is particularly useful in the field of ML, as it increases the number of feature vectors that will be used to train and test the algorithms. In this work, it is argued that the application of this technique will improve the performance of the SVM-based classifier. To test this hypothesis, two feature datasets were created, the first using the first process (i.e., absence of windowing). Whereas the second dataset was created by segmenting the signals using a one-minute window, and computing the features from each segment. Additionally, a 30-second overlap was defined for the windowing process. The duration of one minute was chosen taking into account that it is the minimum signal duration required by the *Neurokit2*, for computing the LF band and HF band features, from the ECG and respiration signals.

4.6 Classification

Regarding the classification task, this work followed the methodology proposed by Raschka [144], which can be summarized as follow:

1. Split data into train/test folds;
2. Preprocessing of the training data;
3. Hyperparameter tuning using cross-validation on the training data;
4. Application of the optimal model on the testing data;
5. Evaluation of the model performance.

Since the model will evaluate performance on two different datasets (see previous section), this work will be comprised of two classification scenarios. These are very similar to each other, so the methodology adopted will be common to both and will be explained in the next sections. Additionally, all steps of the adopted methodology were conducted using the *Scikit* library [68], since it supports SVM multi-class classification, amongst other necessary functionalities.

4.6.1 Data Splitting Into Training and Testing

The first stage of the classification process, referring to the random division of the original dataset into training and testing, is performed in order to determine if an algorithm performs well not only in the dataset used to fit the predictive model (training),

but also the ability to generalize to new observations (test). In the literature, the split of the data is usually done using train/test proportions of 60:40, 70:30 or 80:20, depending on the initial size of the data set, in general the greater the number of observations, the greater the proportion of the initial set used for training the model [144]. To maximize the number of training instances, this work adopted a partitioning of 80:20, for training and testing respectively.

More details on model fitting, as well as on the estimation of the prediction error associated with its application in new observations, will be explored throughout the text. However, it is important to clarify that at all stages of the learning system, only the training set is explored, ensuring, at the end of this process, a more accurate estimate of the predictive performance of the adjusted model when applied to new observations.

The next topic presents the preprocessing step, which explores characteristics of the training dataset in order to prepare it for the learning process.

4.6.2 Data preprocessing

In general, the original data, in its raw form, will not result in the optimal performance of an algorithm. To achieve the best performance it is necessary to consider characteristics related to the size of the data set (number of observations and available predictors), to the response variable (categorical or continuous, balanced or unbalanced) and predictors (continuous variable, categorical variable, different scales, presence of missing values) [144].

4.6.2.1 Feature Scaling

Feature scaling techniques map the features values of a dataset into the same value range. This step is of great importance for ML algorithms that consider the distance between observations, such as the SVM (see section 2.5), because the difference between two observations will be different for non-scaled and scaled data, thus leading to the creation of different models. Thus, feature scaling is performed to avoid features in greater numeric ranges to dominate those in smaller numeric ranges [144].

Standardization, or Z-score normalization, is a common scaling technique and consists of centering variables around a mean (μ) of zero and mapping their values into a mutual range, with a standard deviation (σ) of one, equation 4.1. This method is less affected by outliers in comparison to other scaling techniques such as Scaling Normalization [116, 144], since values are not bounded to a limited range (e.g., [0-1] or [-1, +1]).

$$Z = \frac{x - \mu}{\sigma} \tag{4.1}$$

Hence, for each of the physiological features, extracted from the three biosignals, the mean and standard deviation were calculated, considering all values with the same class label present in the training set. Next, these values were use to standardize each individual feature, using the expression above. Note that the scaling is performed under a population basis, in opposition to a individual basis. This is done considering the inter-variability amongst subjects and to ensure that each physiological feature shares the same numerical range [116, 144].

4.6.2.2 Class Imbalances - SMOTE

In classification tasks, imbalances occur when one or more classes have a low proportion in comparison to other classes. This poses a problem, as most classification algorithms assume a balanced or equal-weight distribution of the data they learn from. In these cases, the algorithms often present high accuracy, but lacks precision, since the leraning process favors the majority class, in detriment of the minority classes.

Of the two datasets created, class imbalances are limited to the dataset generated through windowing. This is due to the fact that there is a difference between the in the amount of signal recorded for each level. Although a effort was made during the elaboration of the experimental protocol to minimize differences in the duration of the recordings made for each of the levels, by manipulating of the number of sequences presented at each level (see section 4.1.3), some differences still persist, as the time necessary to complete each level tends to increase from the easy to the hardest level. This implies that the application of a window with fixed duration results in more instances of the hard class, compared to the medium and easy classes. This is identified as a major limitations of the experimental protocol adopted for this dissertation, since the amount of signal recorded varies according to the speed with which the player performs the task, thus generating class imbalances when windowing is performed.

To solve the class imbalance generated by the windowing segmentation, it was decided to incorporate an additional oversampling step that is applied at the level of the training data. More specifically, the Synthetic Minority Over-sampling Technique (SMOTE) was selected as an appropriate method for creating new synthetic data (i.e., feature vectors) from existing examples of the minority classes. This method developed by Chawla et al. [145] performs oversampling of the positive class, creating new synthetic instances, instead of randomly selecting samples from the dataset. Thus, let X be a training set, with i instances (i.e., feature vectors), and X_M, X_m be the subset of X containing the instances i_M, i_m for the majority and minority classes respectively, under the condition that $i = i_M + i_m$ and $i_M > i_m$. For the subset X_m, select the k-nearest neighbours for each instance x_i of i_m, so that the k neighbours selected have the smallest euclidean distance to the selected vector. Afterwards, randomly select one of the k neighbours selected and

subtract it from x_i. The difference is then multiplied by a random number between 0 and 1 and added to x_i to create a new synthetic instance. This process is repeated for each k neighbour, and looped until the condition $i_M > i_m$ is no longer true [145].

4.6.3 Hyperparamenter Tuning for Model Selection

In ML, most algorithms have one or more parameters that control the complexity (or balance between bias and variance) of the fitted model. Such parameters, called tuning parameters or hyperparameters, have to be directly optimized before fitting the predictive model, as they are not directly estimated by the training data, since there is no analytical formula available to calculate their appropriate value [144]. As addressed in section 2.5, two SVM kernels will be explored in this work, these are the Linear and Radial Basis Function (RBF). Whose selection was based on popularity, behavior and complexity criteria. Moreover, each is characterized by a set of hyperparameters that need to be optimized, to ensure that the best estimator is used during training and testing.

As hyperparameters are related to the complexity and flexibility of a predictive model, inappropriate choices for their values can result, for instance, in overfitting and poor model performance in new observations. Thus, to avoid undesirable optimistic bias and find a robust and reliable indication of the relative performance of competing algorithms, general performance estimation methods such as K-fold Cross-Validation (CV) are employed [146]. In K-fold CV, the dataset is uniformly partitioned, at random, into K folds of similar size. Then, a classifier is trained using $K - 1$ folds, and the performance is determined by testing the classifier in the remaining fold [146].

The selection of the best hyperparameters is made through an exhaustive search of several combinations of parameters, for which the performance is evaluated using K-Fold CV. Thus, the overall performance, for a combination of hyperparameters, is estimated from the scores obtained for each fold. These performance estimates are then compared to select the set of optimal hyperparameter values, that guarantees the selection of the best performing model. In this work, a total of 50 folds (5-Fold CV repeated over 10 random trials) were computed for each kernel, for testing the different combinations over the following set:

- C (regularization) parameter: [10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}]

- γ (Gamma) parameter, specific to the RBF Kernel: [10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}, 10^{6}, 10^{7}]

4.6.4 Application of the optimal model on the testing data

The model determined as optimum is then fitted to the whole training set and is tested on the test set defined in the first step of the classification process (see section 4.6.1). The resulting predictions will be used to compute the confusion matrix. Figure 4.24 illustrates a confusion matrix for a multi-class situation with n classes, and whose fields report on the behavior of a given predictive system, by providing the amount of correct and incorrect classifications for each class. The values of each prediction, defined as c_{ji}, are collected in the confusion matrix C_{ij}, where i corresponds to the label of the true class and j to the label of the estimated class. Generally, with respect to a classification target k, the confusion matrix will report on four types of classification results, from which the first two represent successes, an the last two represent classification errors [147].

- true positives (tp) - the class occurs and is correctly estimated (c_{kk}, where $k = i = j$)
- true negative (tn) - the class doesn't occur and is not estimated ($\sum_{ij \in N \setminus \{k\}} c_{ij}$)
- false positives (fp) - the class doesn't occur and is estimated ($\sum_{i \in N \setminus \{k\}} c_{ik}$)
- false negatives (fn) - the class occurs and is not estimated ($\sum_{i \in N \setminus \{k\}} c_{ki}$)

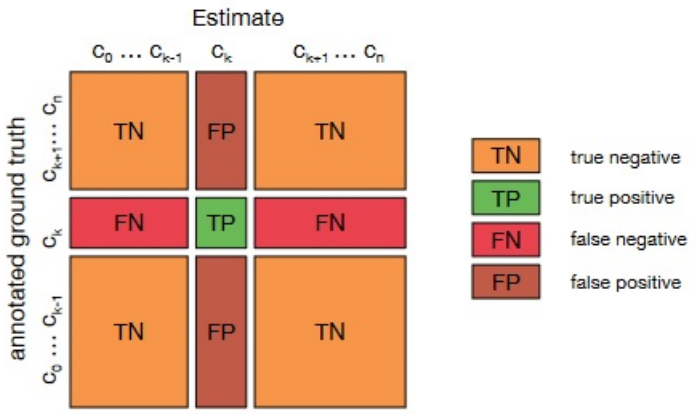

Figure 4.24: Confusion matrix for multi-class classification [147]. The confusion matrix of a classification with n classes

From the confusion matrix, several metrics of interest can be estimated, which will be described in the next section.

4.6.5 Evaluation of performance

There are several evaluation metrics that can be calculated to quantify the performance of a ML model [147]. These metrics are typically organized according to the model's prediction type. This work will only cover those that are commonly used to evaluate

classification problems. These are accuracy, sensitivity, precision and recall. Apart from the accuracy, all the other metrics can be estimated class-wise.

Accuracy provides the amount of correctly classified samples in respect to the total number of predictions made (N). The formula, which provides the general accuracy of a model, is as follows:

$$Accuracy := \frac{\sum_{i=0}^{N} c_{ij}}{\sum_{i=0}^{N} \sum_{j=0}^{N} c_{ij}} \tag{4.2}$$

Sensitivity or **recall** shows the proportion of samples correctly classified as belonging to a class, in respect to all the samples that actually belong to the class. Thus, it provides an indication of the missed positive predictions. Equation 4.3 provides the class-wise sensitivity of the model.

$$Sensitivity := \frac{tp_{class}}{tp_{class} + fn_{class}} \tag{4.3}$$

Precision corresponds to the proportion of samples that were correctly classified as belonging to a class, out of all the samples that were predicted to be of that class. Contrary to sensitivity, this metric provides insight into the number of false positives. It is particularly useful when dealing with unbalanced data. The class-wise precision can be computed from equation 4.4.

$$Precision := \frac{tp_{class}}{tp_{class} + fp_{class}} \tag{4.4}$$

F1-Score provides a measure indicative of the capability of a classifier to predict a given class. It takes into account both precision and sensitivity, by computing the harmonic mean of the two values, as formulated by equation 4.5.

$$F1 - Score := \frac{2 \times tp_{class}}{2 \times tp_{class} + fn_{class} + fp_{class}} \tag{4.5}$$

Although sensitivity, accuracy and f1-score formulas are calculated class-wise, these metrics can be explored in order to assess the overall performance of the model, by averaging the class-wise metric in relation to the total number of classes.

5

Results and Discussion

This chapter will include the most relevant results and their discussion considering the objectives outlined for this dissertation. In this sense, the first section (5.1) will be dedicated to the characterization of the population included in this study. Afterwards (in 5.2), we will proceed to the validation of the TWD videogame proposed for this work, as a legitimate tool to study the effects of difficulty on players emotions. Finally, (in 5.3) the relationship between emotions and difficulty will be studied, through the classification of difficulty using an SVM.

5.1 Sample Characteristics

5.1.1 Demographic Data

A total of 32 individuals volunteered to participate in this study, from whom 34.4% were female. Furthermore, due to previous evidence indicating that performance on the TMT varies significantly with age and education [148], the age of participants was limited to 18-30 years. Thus, the average age of the participants was 22.5 years, with a standard deviation of 1.8. Regarding education, all 32 participants were university students. In addition, no participant reported suffering from any cardio-respiratory disease, as well as hyperhidrosis.

Moreover, participants also reported their experience with videogames and VR, the results are shown in figure 5.1. From the visualization of the bar graph related to the experience with videogames, it is observable that half of the players reported playing on a daily-basis (diary) or several times a week (frequently), while the other half replied that they either played a few times a month, a few times a year or never. As for the experience with VR, about 75% of players reported having already experienced VR, and 43.75% use the technology at least a few times a month.

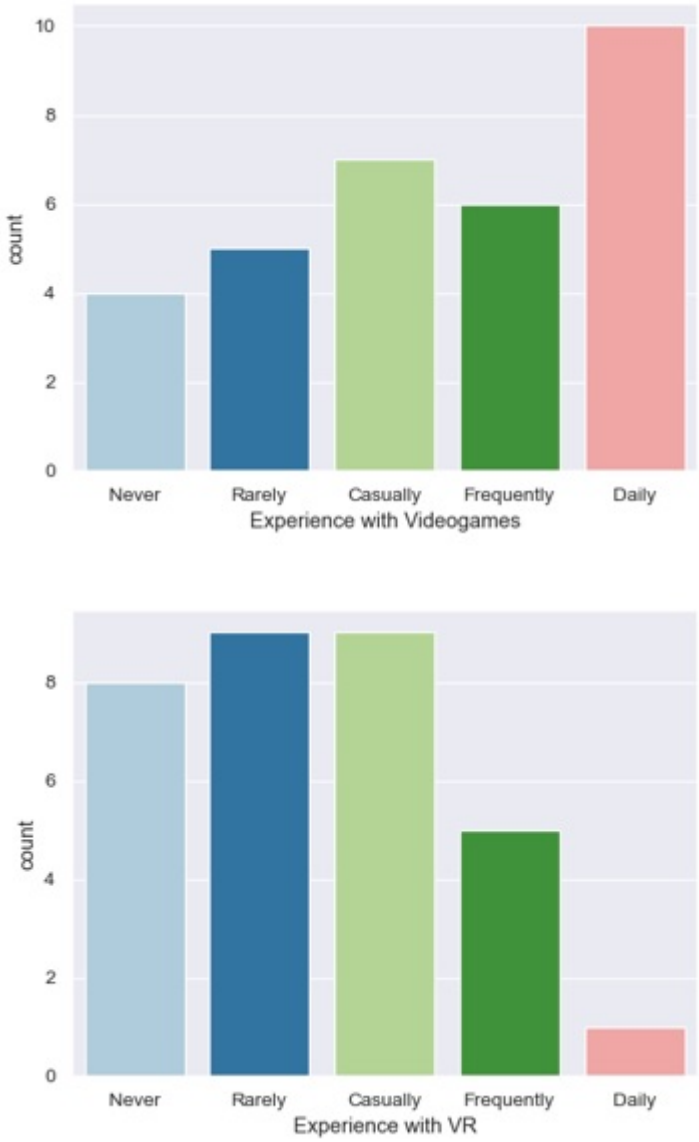

Figure 5.1: Count plots of the 32 participants reports regarding experience with videogames and experience with VR.

5.1.2 Self-assessed Difficulty

At the end of each game level, participants were asked to fill out a series of questionnaires, the first being relative to difficulty. As described in the methodologies (see section 4.2.2), the participant had to report the difficulty experienced during the level, using a scale with the options: *Easy*, *Medium* or *Hard*. In that regard, all players were able to identify the respective difficulty assigned for each level without any prior information. These results contrast with the conclusions obtained for the pilot study (see section 4.1.4), where 75% of the players reported the medium and hard levels as being both of medium difficulty. Furthermore, this results are inline with what was stated for the pilot study

regarding the introduction of a time limit to complete the levels, which would increment the differences in difficulty between the two levels, however, due to differences in the populations used for both experiments, that claim cannot be fully corroborated.

5.1.3 Self-assessed Fatigue

In addition to difficulty, participants were also asked to report their level of physical fatigue at the end of each level (see 4.2.2). This was done in order to identify high levels of physical fatigue, which can directly affect some of the features considered by the SVM-based classifier, for instance heart and breathing rates [136]. Table 5.1 shows the distribution of the reported levels of fatigue for each level.

Table 5.1: Fatigue levels reported for each level. The numbers indicate participant counts. The ranges were defined in accordance with the original questionnaire [136].

Level of Fatigue	Range	Easy	Medium	Hard
No Fatigue At All	0-2	29 (90.625%)	27 (84.375%)	25 (78.125%)
A Little Fatigued	3-4	3 (9.375%)	3 (9.375%)	3 (9.375%)
Moderatly Fatigued	5-7	0	2 (6.25%)	2 (6.25%)
Very Fatigued	8-9	0	0	2 (6.25%)
Total Fatigue	10	0	0	0

5.1.4 Summary Quality Assessment

After analyzing the answers to the questionnaires, as well as checking the quality of the EDA signals, it was concluded that:

- Participants 9, 13, 18, 28 and 32 showed saturated EDA signals.
- Participants 1 and 27 reported very high levels of fatigue (>7).

From theses results, it was decided to exclude these 7 participants from future analysis. Hence, only artifact-free signals were considered for further analysis, thus including data from 25 participants.

5.2 Validation of the Wandering Druid

The Wandering Druid was initially proposed with the objective of studying the effect of difficulty on the players' emotional response, in order to evaluate, through ML, the existence of patterns that allowed establishing a relationship between the two variables. By doing so, this game is intended to validate the application of flow theory to the design of games (as described in section 3.2), as a valuable construct to assess the players emotional state. Thus, depending on the targeted set of emotions, three experimental conditions

regarding challenge were determined: easy condition, medium condition and hard condition. The difficulty associated with each level was pre-set before the experience and fixed for all players (see section 4.1.3). Since experience with videogames and VR can influence the way players perceive difficulty (i.e., an experienced player may be more comfortable playing, compared to a participant who has a poor experience), and there were no guarantees that all participants, who volunteered for this study, would have the same level of experience, each of the levels has been designed with the aim of covering various skill levels. Hence the choice for adopting the TMT as the basis for the game The Wandering Druid, which was made, among other criteria, under the assumption that it provided the necessary accessibility so that the level of experience with VR and videogames, would have a minimal impact on how players learned the game and subsequently performed (see section 4.1.1). Therefore, the following subsections will verify the validity of the adopted game for the purpose of the study, by evaluating the following hypothesis:

- H1: In-game performance is not determined by players' experience with videogames and VR;
- H2: Playing each of the three conditions (difficulty levels) will give rise to different emotional states (proposed in section 3.2);

5.2.1 Testing Hypothesis H1

To assess the validity of hypothesis H1, the performance of players was analysed considering their experience with VR and videogames. In coherence with the pilot study, the performance metric used was time to complete the level. To assess the existence of differences, a Kruskal-Wallis test was used to test whether 2 or more independent groups, of equal or different sample sizes, were statistically different, considering a significance level of $\alpha = 0.05$.

As observed from the bar graph in figure 5.2, groups with different experiences with videogames, share a negligible difference in respect to the performance metric exhibited during the hard difficulty level. A similar behaviour can also be observed for the easy level. Finally, in respect to the medium difficulty, the pattern displayed by the bars does not indicate a clear relationship with respect to each experience group. Note that, as discussed in [14], the medium difficulty level is the most difficult to calibrate, since while the easy and hard levels can be manipulated so that the easy level is very easy, and the hard level is very difficult, the same does not apply to the medium level. Nevertheless, evaluation through the use of a Kruskal-Wallis test, indicates that no statistically significant difference was found between the easy ($H = 7.15,\ p = 0.13$), medium ($H = 7.55,\ p = 0.10$) and hard levels ($H = 6.7,\ p = 0.15$), for the five groups.

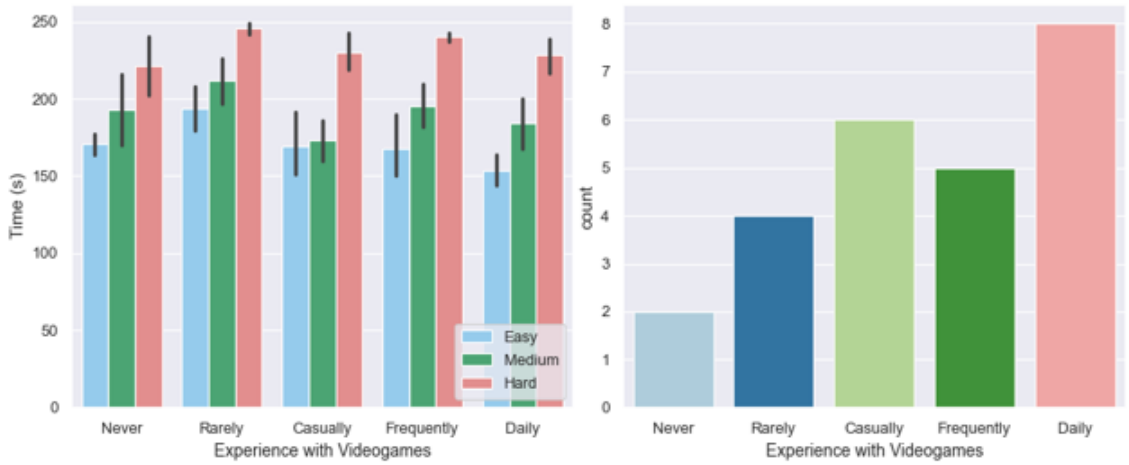

Figure 5.2: Graph of average time (in seconds) to complete each level (the lower the value, the better the performance), grouped by the level of experience with videogames (left). Distribution of the 25 participants by group is displayed on the right.

Regarding the experience with VR, from observation of the bar graph in figure 5.3, a negligible difference is again observed in the performance metric for the high difficulty level. However, it is visually discernible that, in general, the time required to complete the level decreases with the level of experience, for both easy and medium difficulties. This behavior is inline with what was observed in the pilot study and suggests that comfort with VR technology can impact the player's performance, especially at lower difficulty levels. However, the data obtained is insufficient to statistically verify this hypothesis, as the use of Kruskal-Wallis test returns that the differences between the groups, are not statistically significant for the easy ($H = 4.55,\ p = 0.33$), medium ($H = 4.25,\ p = 0.317$) and hard levels ($H = 1.6,\ p = 0.81$).

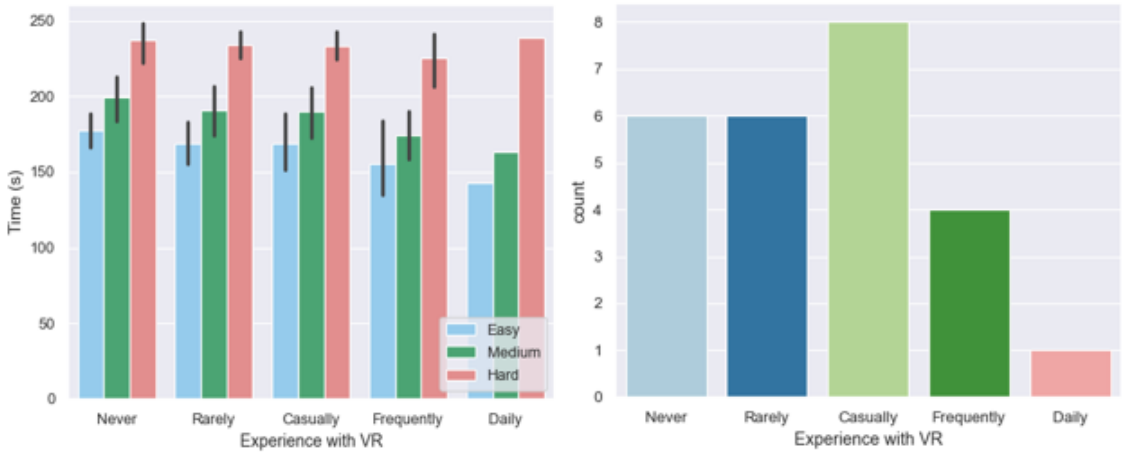

Figure 5.3: Graph of average time to complete each level (the lower the value, the better the performance), grouped by the level of experience with VR (left). Distribution of the 25 participants by group is displayed on the right.

5.2.2 Testing Hypothesis H2

5.2.2.1 Preliminary Analysis of Self-Assessed Ratings

As stated in section 3.2, each difficulty level targeted a specific set of emotions, for instance, the easy level aimed to elicit negative and calm emotional states (low valence - low arousal), whereas the medium level targeted positive and aroused emotional states (high valence - high arousal). Finally, with the increase in difficulty, the purpose of the hard level, was to evoke negative and aroused emotional states, (low valence - high arousal). In order to assess the success of each level in eliciting the desired emotions, the results of the SAM questionnaire are hereafter explored, in terms of the valence and arousal ratings. By taking advantage of dimensional nature of the emotion annotation items, the participants self-reports were categorized into 4 groups: positive valence (>5) and low arousal (<5) (PVLA), positive valence and high arousal (>5) (PVHA), negative valence (<5) and low arousal (NVLA), and negative valence (<5) and high arousal (NVHA).

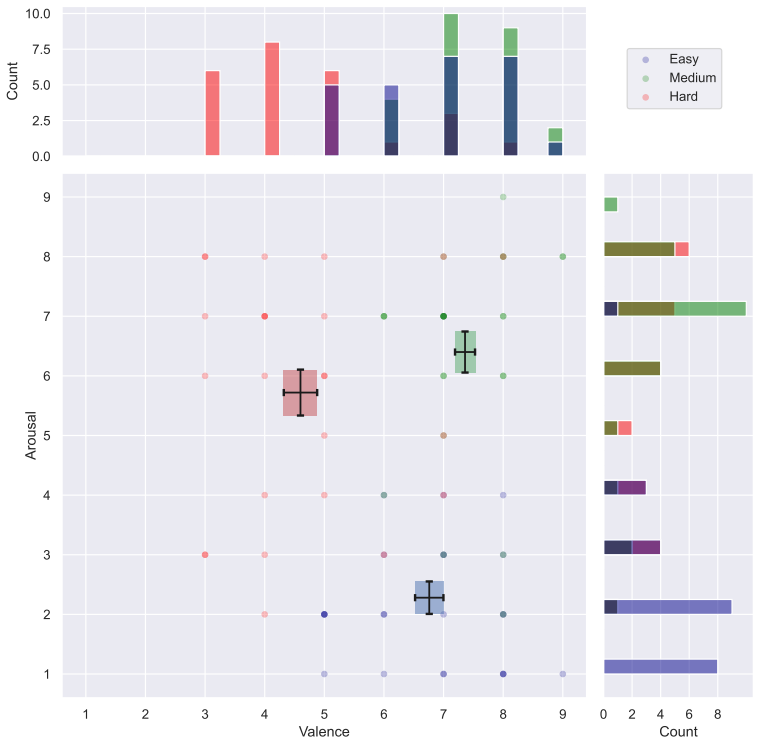

Figure 5.4: SAM Self-Assessed Ratings of 25 participants, mapped to the valence-arousal space. The ratings for the easy level are mostly concentrated in the lower right region of the space $C_{easy}(M_V = 6.76 \pm 0.24,\ M_A = 2.28 \pm 0.27)$, for the medium level are in the upper right region of the space $C_{medium}(M_V = 7.36 \pm 0.17,\ M_A = 6.4 \pm 0.34)$, and for the hard level the ratings extend to both the upper and lower left regions $C_{hard}(M_V = 4.6 \pm 0.28,\ M_A = 5.72 \pm 0.38)$.

Figure 5.4 is the result of mapping the self assessed ratings of the 25 participants to the valence-arousal space. Each point corresponds to the ratings given for the easy, medium or hard levels, i.e., blue, green and red points respectively. Each level is described by a box, whose center (C_{level}) is determined by computing the mean of the reported valence (M_V) and arousal (M_A) values. Moreover, the dimensions of each box are bounded by the standard deviation estimated using bootstrap on the ratings given for each dimension. A closer look at the figure, reveals that, although there are three discernible regions within the valence-arousal space, associated with each of the difficulty levels, it is challenging to define the limits of separation between the three clusters of points, due to the presence of outliers in the data, for instance, for the easy level, one of the players stands out from the rest by having reported an arousal level equal to 7. In a similar manner, although most ratings for the medium level are located in the upper right region of the valence and arousal space (PVHA), some of the reported responses extend to the lower right regions (PVLA). These behaviors become more evident, once the data is plotted through a boxplot, as can be seen in figure 5.5.

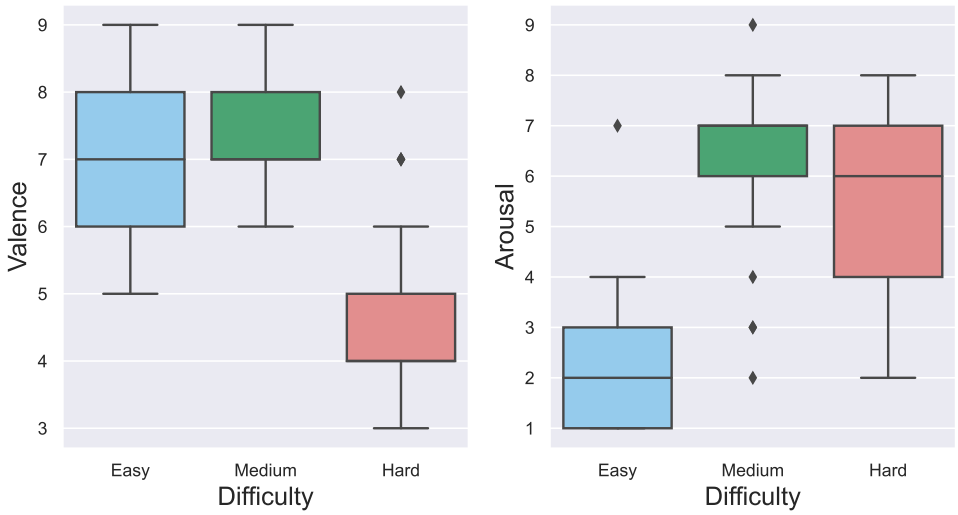

Figure 5.5: Boxplots of the self-assessed ratings of 25 participants. Several outliers can be visually identified for both valence and arousal ratings.

Hence, analysis of the data associated with each of the boxes, leads to the identification of 7 outliers, corresponding to participants 2, 7, 8, 10, 12, 15 and 23. Removing these participants from the dataset and re-plotting of the self-assessed ratings of the remaining 18 participants, to the valence-arousal space, yields significantly better results. As can be seen in figure 5.6, the three clusters of points are now easily discernible, and can be summarized as follows:

- The ratings reported for the easy difficulty level are concentrated in the lower

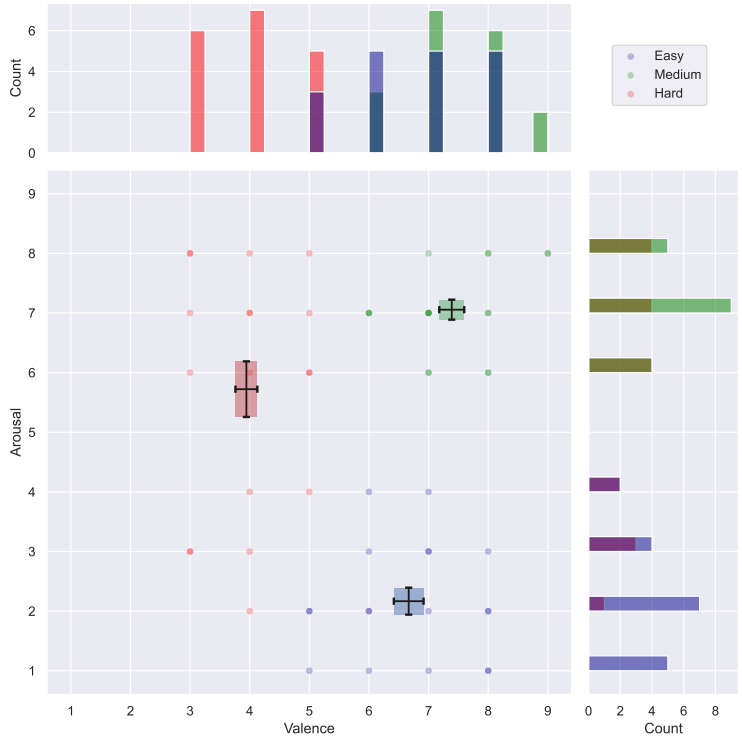

Figure 5.6: SAM Self-Assessed Ratings of 18 participants, mapped to the valence-arousal space. The ratings for the easy level are mostly concentrated in the lower right region of the space $C_{easy}(M_V = 6.67 \pm 0.25,\ M_A = 2.17 \pm 0.23)$, for the medium level are in the upper right region of the space $C_{medium}(M_V = 7.39 \pm 0.21,\ M_A = 7.06 \pm 0.17)$, and for the hard level the ratings extend to both the upper and lower left regions $C_{hard}(M_V = 3.94 \pm 0.18,\ M_A 5.72 \pm 0.47)$.

right region of the valence and arousal space (PVLA). The box $C_{easy}(M_V = 6.67 \pm 0.25,\ M_A = 2.17 \pm 0.23)$ indicates that both dimensions share similar variability.

- The upper right region of the valence and arousal space (PVHA), is where most of the ratings reported for the medium difficulty level are concentrated. In this case, the box $C_{medium}(M_V = 7.39 \pm 0.21,\ M_A = 7.06 \pm 0.17)$ shows a similar variability amongst the reported valence and arousal ratings.

- The ratings reported for the hard difficulty level are divided between the upper left (NVHA), and the lower left (NVLA) regions of the valence-arousal space. As a consequence, the center of the cluster $C_{hard}(M_V = 3.94 \pm 0.18,\ M_A = 5.72 \pm 0.47)$ is close to the neutral value of arousal (5), and presents a higher variability, in comparison to the valence ratings.

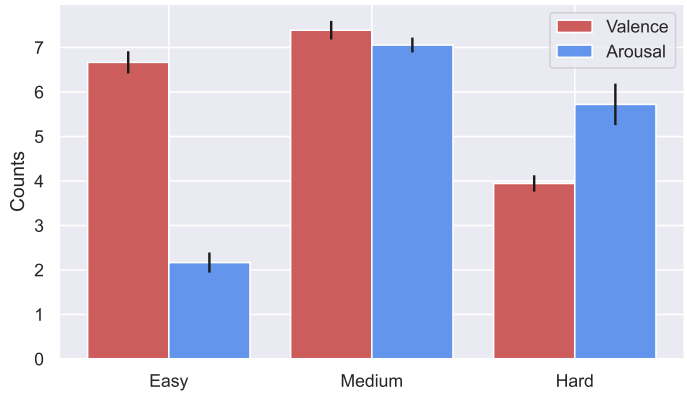

Figure 5.7: Bar graph of the self-assessed ratings of 18 participants for valence and arousal, grouped by level.

Furthermore, assessment of how each dimension varies with difficulty can be done by referring to figure 5.7. As can be seen, mean valence varies little between the easy and medium levels, however arousal increases by nearly 5 points, which hypothesises (H_a) that playing the medium level was more arousing then playing the easy level. Additionally, when comparing the medium and hard levels, it is observed that the mean value for both dimensions decreases. This raises the hypothesis (H_b) that playing the medium level is a more positive and arousing experience, than playing the hard level. Finally, regarding the comparison between the easy and hard levels, there is a clear difference between both dimensions, which brings up hypothesis (H_c), that playing the easy level is a more positive, but less arousing experience in comparison to the hard level. These three hypothesis will be hereafter evaluated.

5.2.3 Statistical Analysis of Self-Assessed Ratings

The evaluation of the statistical significance of hypotheses H_a, H_b and H_c, required the statistical analysis of the data, to determine the most appropriate statistical test. Thus, six data groups (corresponding to the 3 levels for each emotional dimension) were evaluated for skewness and kurtosis. The results, presented in table 5.2, indicated normality, considering the reference values defined in [131]. Hence, it was decided to apply a pairwise T-test to evaluate the existence of significant differences between pairs of groups. Statistical significance was set to $\alpha = 0.05$ and corrected for multiple comparisons, using a Bonferroni correction [134]. The results are summarized as follows:

- H_a: The T-test revealed that the average of arousal for the medium level was significantly higher compared to the easy level ($T = -16.84$, $p_{corrected} << 0.016$), thus verifying the hypothesis that the medium level was more arousing than the easy

Table 5.2: Verification of the assumption of normality, a necessary condition for the application of parametric statistical tests. Asymmetry (skewness) and kurtosis were calculated.

Level	Skewness	Kurtosis
Easy	-0.17	2.39
Medium	0.11	2.09
Hard	0.11	1.85

one. In fact, application of the same test to the valence data results in a similar conclusion ($T = -2.72$, $p_{corrected} = 0.022$).

- H_b: It was identified that the means of arousal and valence were both significantly higher than those of the hard level (valence, $T = 12.17$, $p_{corrected} << 0.016$; arousal, ($T = 3.23$, $p_{corrected} = 0.007$)), thus validating that the medium level provided simultaneously a more arousing and positive experience than the hard level.

- H_c: The comparison of the arousal means for the hard and easy levels revealed that the first was significantly higher than the second ($T = 7.21$, $p_{corrected} << 0.0016$), which verifies that the easy level was less arousing than the hard level. The results of the same test applied to the valence means indicate that the values for the easy level were significantly higher than those for the hard level ($T = 7.38$, $p_{corrected} << 0.016$), thus validating the hypothesis that the easy level gave rise to a more positive experience compared to the hard level.

5.2.4 Summary Results and Considerations

The results for hypothesis H1 demonstrated that increasing experience with either videogames or VR didn't affect how players performed, thus validating the hypothesis that the game was simple enough to successfully minimize effects associated with different levels of gaming experience, or easiness with the VR equipment. However, it is important to note that an increase in the population used to perform these tests may result in different conclusions. Furthermore, these tests did not take into account the non-uniformity of the distributions relative to each group, associated for instance with the number of samples, or gender representation. This limitations should be taken in consideration for future protocols.

As for hypothesis H2, from the analysis of the SAM questionnaires, it was found that the game was effective in eliciting different emotions for each of the levels of difficulty, however these do not necessarily correspond to the emotions initially targeted. For instance, the easy level was designed with the purpose of eliciting emotions in the lower left region (NVLA) of the valence-arousal space, since it's the region associated with more negative and less arousing discrete emotions, such as boredom. What was found is that most of the players' responses are located in the lower right region (PVLA), which is associated with less arousing, but more positive discrete emotions, which suggests that

players were in a relaxed state while playing this level. On the other hand, the ratings for the hard level indicate that the generality of players considered the overall experience as being negative (valence < 5), however the same can't be said for the ratings of arousal, that vary between the upper region (NVHA) of the valence-arousal space, normally associated with discrete emotions such as anxiety and frustration, and the lower region (NVLA), usually associated with discrete emotions like boredom. This contradicts the emotional elicitation objectives initially proposed for the hard level, whose target was limited to the right upper region of the space (NVHA). Although it is not possible to say with full certainty the reason why the proposed elicitation protocol failed, one can point out some potential causes. For the easy level, the results may be associated with the fact that prior to the experiment no player had any knowledge about the game, so during the easy level the player was focused in mastering the basic mechanics of the game. In this case, to potentially elicit the desired emotions, the duration of the level would have to be extended with more sequences to complete. As for the hard level, the difference in results can be associated on one hand with the fact that all levels are played in the same order, i.e. playing the medium level followed by the hard level, this in turn can influence players' perspective of the harder level. On the other hand, individual traits of the players may play a important role, namely the level of competitiveness, as different characteristics of competitiveness can lead to different changes in the motivation and mood of the players exposed to similar gaming scenarios [149]. To compensate for this difference, a monetary reward could be incorporated in order to motivate more players. Nevertheless, these claims require a more thorough investigation in order to verify their validity and usefulness. Only the ratings for the medium level coincide with those initially proposed as target emotional states for the medium difficulty, that is, the upper right region of the valence-arousal space (PVHA). Although the emotions elicited did not correspond to those initially proposed, particularly for the easy and hard levels, reflecting on the results it makes sense that most players felt relaxed during the easy experience; moreover, with an increase in challenge, players seem to be more attentive and enjoying the experience, suggesting they are in state of flow. With the hard level, players still experience high arousal, however confusion or frustration begin to develop. As such, the next section will explore whether these emotional states translate into distinct physiological responses in players; and whether these differences are consistent among players to allow their prediction using an SVM-based classifier.

5.3 Classification

In this section, the results related to the classification task will be presented and discussed. As initially proposed, it will consist of two classification scenarios associated with two datasets generated using different methodologies - with (B) or without (A) windowing applied (see section 4.6). Furthermore, considering the results of the previous analysis, only data from 18 participants was considered for this evaluation. The steps for

each classification scenario are similar and discussed hereafter.

5.3.1 General Workflow

Each dataset was partitioned using a split of 80% for training and 20% for testing. Afterwards, the selection of the set of hyperparameters that allowed maximizing the performance of the Linear and RBF SVM classifiers was carried out on the training set. To this end, the K-Fold CV procedure was used to perform an exhaustive search of the various combinations of parameters proposed in section 4.6. The performance of each set of tuning parameters was performed through a 5-fold CV, repeated 10 times, making it a total of 50 folds tested for each set of hyperparameters. In the case of the dataset A, data was standardized for each fold, whereas in the case of dataset B, an additional step of oversampling of the minority classes was employed. The results are summarized in table 5.3.

Table 5.3: The best performing set of hyperparameters for the Linear and RBF SVMs determined through exhaustive search over dataset A and B. The γ (gamma) parameter is only computed for the RBF kernel.

	Dataset A		**Dataset B**	
Parameter	**Linear SVM**	**RBF SVM**	**Linear SVM**	**RBF SVM**
C	0.1	10^4	1	10
γ (gamma)	-	10^{-6}	-	0.01

5.3.2 Classifier Performance - Dataset A

The optimal linear SVM and RBF models determined for dataset A were both applied to the test set and the results of the confusion matrix are presented below (figure 5.8), along with the metrics calculated from the data (tables 5.4 and 5.5). As can be seen, the accuracy of both classifiers was sub-optimal, since both performed below the uniform random accuracy of classifying 3 classes (33,33%), which means that none of the models could find patterns that allowed the correct classification of the three levels of difficulty experienced during the elicitation protocol.

Table 5.4: Linear SVM classification results obtained for dataset A. The value of each metric is presented for each of the classes, along with the total average (weighted by the number of elements in each class).

Class	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Easy	-	20%	33%	25%
Medium	-	0%	0%	0%
Hard	-	67%	50%	57%
Average	27%	30%	27%	28%

Table 5.5: RBF SVM classification results obtained for dataset A. The value of each metric is presented for each of the classes, along with the total average (weighted by the number of elements in each class).

Class	Accuracy	Precision	Recall	F1-Score
Easy	-	29%	67%	40%
Medium	-	0%	0%	0%
Hard	-	0%	0%	0%
Average	18%	8%	18%	11%

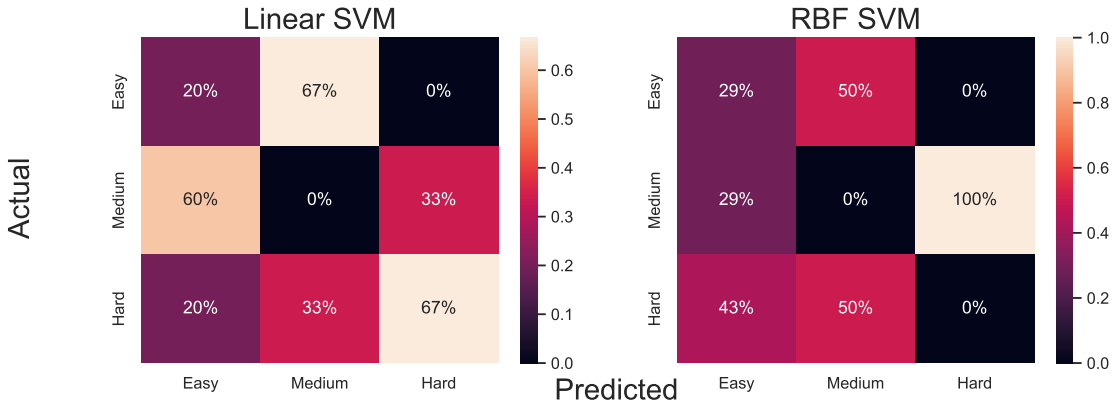

Figure 5.8: Confusion Matrix's obtained for the Linear SVM and RBF SVM using dataset A. Values were normalized by the predicted values, so that the diagonal of the matrix provides information about the precision of both models.

Both models lack generalization ability, which is particularly true for the RBF case, given the high C (regularisation) value, which is indicative of over fitting. Additionally, compared to linear SVMs, RBF SVMs are more susceptible to hyperparameter values, so it is normal to potentially exhibit worse results compared to Linear SVM. As for why both models underperformed, this behavior can be related, for instance with the high dimensionality, i.e., the number of features (63) and low sample size (58) context associated with both classification tasks, as described in [150]. Additionally, using a large temporal window (i.e., the full duration of the acquisition) to compute the features, may lack the granularity necessary to detect high-intensity and low-duration physiological events. It is thus believed that the existence of a larger dataset would yield significantly better results, as will be discussed next.

5.3.3 Classifier Performance - Dataset B

The results of testing the optimal models with new data from dataset B are presented in tables 5.6 and 5.7, along with the respective confusion matrices 5.9. As can be seen, both models performed significantly better, compared to the results obtained for dataset A. Furthermore, comparison of the two kernels, indicates that the RBF-based SVM performed

better with respect to all metrics. From the analysis of the F1-Scores, it is concluded that both models had a more difficult time predicting the medium difficulty class, compared to the remaining classes. This is in line with the results found in the literature [14], and highlights the variability between players, regarding the notion of flow or engagement.

Table 5.6: Linear SVM classification results obtained for dataset B. The value of each metric is presented for each of the classes, along with the total average (weighted by the number of elements in each class).

Class	Accuracy	Precision	Recall	F1-Score
Easy	-	71%	77%	74%
Medium	-	61%	48%	54%
Hard	-	62%	70%	65%
Average	65%	64%	65%	64%

Table 5.7: RBF SVM classification results obtained for dataset B. The value of each metric is presented for each of the classes, along with the total average (weighted by the number of elements in each class).

Class	Accuracy	Precision	Recall	F1-Score
Easy	-	68%	77%	72%
Medium	-	60%	65%	63%
Hard	-	78%	61%	68%
Average	68%	69%	68%	68%

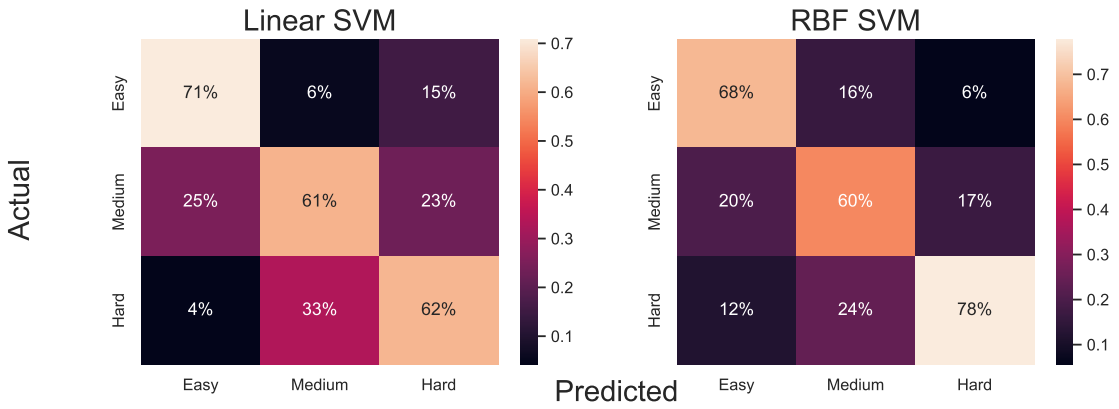

Figure 5.9: Confusion Matrix's obtained for the Linear SVM and RBF SVM using dataset B. Values were normalized by the predicted values, so that the diagonal of the matrix provides information about the precision of both models.

For reference, the precision of both models was also compared with and without SMOTE as indicated in figure 5.10. As initially proposed, the oversampling of the minority classes increased the overall precision for both models, although the effect was most

noticeable for the RBF SVM. Note that the use of SMOTE can affect the choice of the best set of hyperparameters, for this reason it was necessary to perform an exhaustive search again to tune the hyperparameters without SMOTE, however the search results were the same for both conditions.

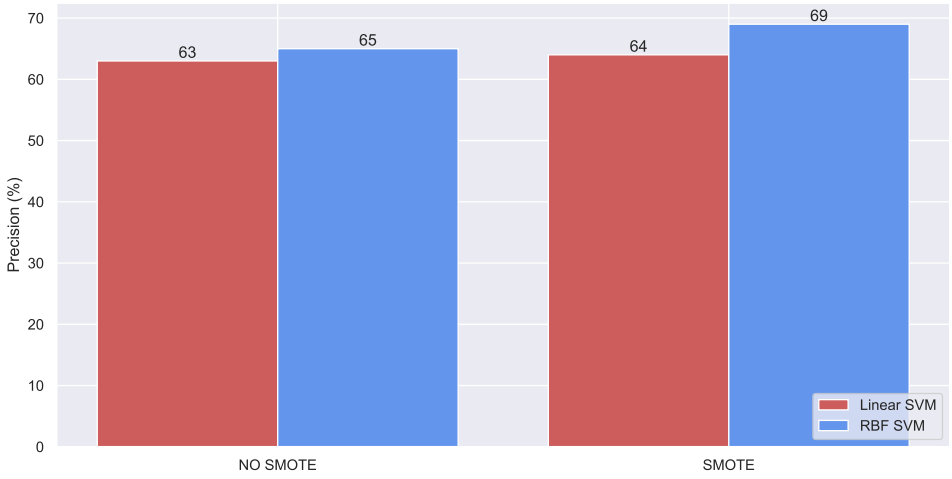

Figure 5.10: Bar graph of the precision's of both classifiers in a context of unbalanced ("NO SMOTE") and balanced classes ("SMOTE").

5.3.4 Summary Results and Considerations

The classification task results demonstrate that for dataset A, without windowing applied, both classification models performed significantly worse than the results found in the literature. For instance, in the work by Chanel et al. [14], the results, without any kind of windowing, point to an overall accuracy of 59%, using peripheral signals. Although not fully explainable, one can point at least two potential reasons to explain it; firstly in the compared study, the total number of features employed for the classification task was less than the number of samples used. This is not the case for the condition of dataset A, which is performed in a context of high dimensionality and low sample count. This can result in a problem known as the "curse of dimensionality", which is caused by the exponential increase in volume associated with adding extra dimensions to Euclidean space, having that in high dimensions all feasible training samples will sparsely populate the input space [150]. To minimize the effects associated with this problem, dimensionality reduction techniques can be employed [116]. The other hypothesis is related to the type and importance of the features used as input for the predictor, for example, in the compared work, the set of features were subjected to a selection process in order to reduce the dimensionality of the problem, but also to eliminate features that mainly introduced

noise, and contributed little to the classification. This can significantly increase the performance of the classifier, and could be a good approach to further explore. Additionally, the way features are extracted can also have a significant impact on the performance of the classifier, i.e., during exposure to a complex stimulus, such as that of a videogame, an individual will not be under the effect of a continuous affective state. In fact, it is expected that parts of the signal will lack relevant information to the predictive task, so using a very large time window can make it difficult to detect the relevant physiological events, that arise as a consequence of the player experiencing the target emotion.

Therefore, with dataset B, this work sought to minimize some of the limitations mentioned above, through the use of windowing and oversampling (SMOTE) techniques, which resulted in a larger dataset on which to train the models. As a result, the performance of both classifiers was significantly higher, presenting a higher predictive ability than that of [14], using only peripheral information. In general, the RBF kernel-based SVM performed better than the linear SVM, however both models share a lower detection rate of the medium difficulty, when compared to the other classes. This can be mainly attributed to the high variability between players, i.e. different players may have different approaches to the game, for example, some may prefer it when the difficulty slightly exceeds their abilities, while other may enjoy a more balanced or easy task. Thus, as the stimulus is the same (level was the same for all players), this can trigger different intensities of emotion, or even different emotions, which will later translate into different physiological patterns. This granularity was not considered in the analysis performed, since both dimensions of the emotion rating (i.e., valence and arousal) were downscaled and evaluated in accordance to a high(>5) or low (<5) notion, and did not take into account potential differences between players who reported, for example, an arousal 6 and others who reported an 8. This may have compromised the ability of the classifier to predict the medium difficulty class, since a higher variability between the samples is expected in this case. Nevertheless, the results allow validating the hypothesis that varying the difficulty of a game can elicit distinct emotional states, and that these can be automatically detected using an SVM-based classifier.

6

Conclusion and Future Work

6.1 Conclusion

This dissertation investigated the autonomous assessment of difficulty from physiological data using a SVM-based classifier. As part of the experimental protocol adopted for this work, it was necessary to design and implement a VR videogame that, through different levels of difficulty, could elicit different emotions. The proposed solution was the Wandering Druid, a puzzle game, characterized by three parts, each offering a distinct level of challenge. The proposed game was then used in an experiment where the physiological signals and the self-report data of 32 participants were gathered, for each level of difficulty played. Afterwards, two types of analysis were performed on the collected data. On one hand, a statistical analysis of the self-assessed ratings was conducted with the intent of validating the proposed game in terms of accessibility and emotion elicitation effectiveness. On the other hand, classification was performed with the objective of verifying if it was possible detect difficulty taking into account the physiological patterns associated with each elicited set of emotions.

The results obtained from the statistical analysis of the self-reports showed that different levels of experience with either VR or videogames didn't have a measurable impact on how players performed during the three levels. Additionally, the self-assessed emotional ratings, indicated that playing the Wandering Druid at different difficulty levels gave rise to different emotional states. The easy level was related to a state of positive valence, and low arousal. In comparison, the medium level was regarded as a more arousing and positive experience that the easy level. Finally, for the hard level, participants reported the experience as negative and less arousing than the medium level, although compared to the easy level it was more exciting. The results indicate that despite the easy and hard levels of the game not being able to elicit the emotions initially intended, the protocol adopted was successful in eliciting different emotions for each level, thus validating the usability of the Wandering Druid as a medium to explore the effects of difficulty over players' emotions.

The automatic detection of the three levels of difficulty using the peripheral biosignals recorded during each of the conditions was analyzed for different SVM-based classifiers, and time windows. The results obtained indicate that the RBF-based SVM (F1-score = 68%) is more suitable for the prediction of the three levels of difficulty, compared to the Linear SVM (F1-score = 64%). Moreover, analysis of the performance of the classifiers using features computed in different temporal windows shows that in the absence of windowing, both classifiers underperformed, which is explained by the high dimensionality and low sample size context of the dataset used, whilst the results for the segmentation using 1-minute windows are in accordance with the literature, and can be considered promising. Overall, the results verified the hypothesis that the physiological data, associated with the emotional states elicited during gameplay of the three conditions, can be used to predict game difficulty.

6.2 Future Work

Future work will include changes to both the elicitation protocol and classification task. Thus, regarding the experimental protocol, these can be summarized as follows:

- The database has to be extended, while keeping in mind the balance between genders, as well as experience with both VR and videogames.
- The easy difficulty level will have to be redesigned in order to extend its duration. This can be done by including more sequences to complete. Additionally, to avoid long acquisition sessions that would make the participants tired and unable to feel the target emotions, acquisitions could be divided into 3 separate sessions.
- To compensate for the variability of emotions reported for the hard difficulty level, a prize can be included as a way of motivating more participants to invest more in the game.
- Questionnaires should also assess players' gaming experience, in addition to emotions reported using the SAM. To this end, we propose the integration of questionnaires such as the Game Experience Questionnaire (GEQ) [151], to assess flow, immersion and presence.

Moreover, future changes and objectives related to the classification task, are hereafter summarized:

- Future work should focus on testing the classification performance of other supervised learning methods.
- Methods for dimensionality reduction or feature selection should be integrated into the classification workflow and their effect explored on the overall performance of the predictive task.

Bibliography

[1] S. Thushara and S. Veni. "A multimodal emotion recognition system from video". In: *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*. IEEE, 2016.

[2] P. Ekman, E. R. Sorenson, and W. V. Friesen. "Pan-Cultural Elements in Facial Displays of Emotion". In: *Science* 164.3875 (1969), pp. 86–88.

[3] J. A. Russell. "A circumplex model of affect". In: *J. Pers. Soc. Psychol.* 39.6 (1980), pp. 1161–1178.

[4] J. May. "Human–Computer Interaction". In: *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, 2001, pp. 7031–7035.

[5] R. W. Picard. *Affective Computing*. Cambridge, MA: MIT Press, 1997.

[6] H. Monkaresi et al. "Automated detection of engagement using video-based estimation of facial expressions and heart rate". In: *IEEE Trans. Affect. Comput.* 8.1 (2017), pp. 15–28.

[7] S. Saha et al. "A study on emotion recognition from body gestures using Kinect sensor". In: *2014 International Conference on Communication and Signal Processing*. IEEE, 2014.

[8] R. Calvo et al. "Emotion Modeling for Social Robots". In: *The Oxford Handbook of Affective Computing*. Oxford University Press, 2015.

[9] P. Rani et al. "An empirical study of machine learning techniques for affect recognition in human-robot interaction". In: *Pattern Anal. Appl.* 9.1 (2006), pp. 58–69.

[10] E. Yadegaridehkordi et al. "Affective computing in education: A systematic review and future research". In: *Comput. Educ.* 142.103649 (2019), p. 103649.

[11] M. Pezzera and N. A. Borghese. "Dynamic difficulty adjustment in exer-games for rehabilitation: a mixed approach". In: *2020 IEEE 8th International Conference on Serious Games and Applications for Health (SeGAH)*. IEEE, 2020.

[12] R. Calvo et al. "Emotion in Games". In: *The Oxford Handbook of Affective Computing*. Oxford University Press, 2015.

[13] J. Nakamura and M. Csikszentmihalyi. "The concept of flow". In: *Flow and the Foundations of Positive Psychology*. Dordrecht: Springer Netherlands, 2014, pp. 239–263.

[14] G. Chanel et al. "Emotion assessment from physiological signals for adaptation of game difficulty". In: *IEEE Trans. Syst. Man Cybern. A Syst. Hum.* 41.6 (2011), pp. 1052–1063.

[15] B. Yang and M. Lugger. "Emotion recognition from speech signals using new harmony features". en. In: *Signal Processing* 90.5 (2010), pp. 1415–1423.

[16] I. B. Mauss and M. D. Robinson. "Measures of emotion: A review". In: *Cogn. Emot.* 23.2 (2009), pp. 209–237.

[17] D. Goleman. *Emotional Intelligence*. New York, NY: Bantam Books, 2014.

[18] B. Reeves and C. Nass. *The Media equation: how people treat computers, television, and new media*. Cambridge University Press, 1997.

[19] S. Poria et al. "A Review of Affective Computing". In: *Inf. Fusion* 37.C (2017), pp. 98–125. ISSN: 1566-2535.

[20] M. Egger, M. Ley, and S. Hanke. "Emotion recognition from physiological signal analysis: A review". en. In: *Electron. Notes Theor. Comput. Sci.* 343 (2019), pp. 35–55.

[21] B. Bontchev. "Adaptation in affective video games: A literature review". In: *Cybern. Inf. Technol.* 16.3 (2016), pp. 3–34.

[22] S. Fairclough and K. Gilleade. "Construction of the Biocybernetic Loop: A Case Study". In: *Proceedings of the 14th ACM International Conference on Multimodal Interaction*. New York, NY, USA: Association for Computing Machinery, 2012, pp. 571–578.

[23] S. P S and M. G S. "Emotion Models: A Review". In: *International Journal of Control Theory and Applications* 10 (Jan. 2017), pp. 651–657.

[24] J. Mizgajski and M. Morzy. "Affective recommender systems in online news industry: how emotions influence reading choices". In: *User Modeling and User-Adapted Interaction* 29 (2019).

[25] T. Eerola and J. K. Vuoskoski. "A comparison of the discrete and dimensional models of emotion in music". en. In: *Psychol. Music* 39.1 (2011), pp. 18–49.

[26] A. Ortony and T. J. Turner. "Whats Basic About Basic Emotions". In: *Psychological Review* 97.3 (1990), pp. 315–331.

[27] M. T. o. Mendl. "On the evolution and optimality of mood states". In: *Behavioral Sciences* 3.3 (2013), pp. 501–521. ISSN: 2076-328X.

[28] O. Bălan et al. "Emotion classification based on biophysical signals and machine learning techniques". In: *Symmetry* 12.1 (2020), pp. 1–22. ISSN: 20738994.

[29] J. A. Russell and A. Mehrabian. "Evidence for a three-factor theory of emotions". In: *J. Res. Pers.* 11.3 (1977), pp. 273–294.

[30] S. Hamann. "Mapping discrete and dimensional emotions onto the brain: controversies and consensus". In: *Trends in Cognitive Sciences* 16.9 (2012), pp. 458–466. ISSN: 1364-6613.

[31] C. Peter and A. Herbon. "Emotion representation and physiology assignments in digital systems". In: *Interacting with Computers* 18.2 (2006), pp. 139–170. ISSN: 0953-5438.

[32] A. Mehrabian. "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament". In: *Current Psychology* 14.4 (1996), pp. 261–292.

[33] A. Kelava et al. "A new approach for the quantification of synchrony of multivariate non-stationary psychophysiological variables during emotion eliciting stimuli". en. In: *Front. Psychol.* 5 (2014), p. 1507.

[34] H. A. Elfenbein and N. Ambady. "On the universality and cultural specificity of emotion recognition: A meta-analysis". In: *Psychol. Bull.* 128.2 (2002), pp. 203–235.

[35] H. A. Osman and T. H. Falk. "Multimodal affect recognition: Current approaches and challenges". In: *Biological Signals and Images*. Ed. by S. A. Hosseini. 2017.

[36] P. Ekman, R. W. Levenson, and W. V. Friesen. "Autonomic Nervous-System Activity Distinguishes among Emotions". In: *Science* 221.4616 (1983), pp. 1208–1210.

[37] S. Jerritta et al. "Physiological signals based human emotion Recognition: a review". In: *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*. IEEE, 2011.

[38] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas. "Human emotion recognition: Review of sensors and methods". en. In: *Sensors (Basel)* 20.3 (2020), p. 592.

[39] J. Zhang et al. "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review". In: *Inf. Fusion* 59 (2020), pp. 103–126.

[40] Y. Gu, K.-J. Wong, and S.-L. Tan. "Analysis of physiological responses from multiple subjects for emotion recognition". In: *2012 IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 2012.

[41] L. Santamaria-Granados et al. "Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS)". In: *IEEE Access* 7 (2019), pp. 57–67.

[42] B. Zhong et al. "Emotion recognition with facial expressions and physiological signals". In: *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2017.

[43] O. AlZoubi et al. "Detecting naturalistic expression of emotions using physiological signals while playing video games". en. In: *J. Ambient Intell. Humaniz. Comput.* (2021).

[44] M. Maier et al. "DeepFlow: Detecting optimal user experience from physiological data using deep neural networks". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. California: International Joint Conferences on Artificial Intelligence Organization, 2019.

[45] Y. Dai et al. "Reputation-driven multimodal emotion recognition in wearable biosensor network". In: *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*. IEEE, 2015.

[46] A. Haag et al. "Emotion recognition using bio-sensors: First steps towards an automatic system". In: *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 36–48.

[47] A. Greco, G. Valenza, and E. P. Scilingo. *Advances in Electrodermal Activity Processing with Applications for Mental Health: From Heuristic Methods to Convex Optimization*. Springer International Publishing.

[48] M. van Dooren, J. J. G. G.-J. de Vries, and J. H. Janssen. "Emotional sweating across the body: comparing 16 different skin conductance measurement locations". en. In: *Physiol. Behav.* 106.2 (2012), pp. 298–304.

[49] H. Sequeira et al. "Electrical autonomic correlates of emotion". en. In: *Int. J. Psychophysiol.* 71.1 (2009), pp. 50–56.

[50] S. H. Fairclough. "Fundamentals of physiological computing". In: *Interacting with Computers* 21.1-2 (2009), pp. 133–145.

[51] L. Petrescu et al. "Integrating biosignals measurement in virtual reality environments for anxiety detection". In: *Sensors (Switzerland)* 20.24 (2020), pp. 1–32.

[52] W. Boucsein et al. "Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures. Publication recommendations for electrodermal measurements". In: *Psychophysiology* 49 (2012), pp. 1017–1034.

[53] S. Taylor et al. "Automatic identification of artifacts in electrodermal activity data". en. In: *Annu Int Conf IEEE Eng Med Biol Soc* 2015 (2015), pp. 1934–1937.

[54] P. J. Bota et al. "A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals". In: *IEEE Access* 7 (2019), pp. 140990–141020.

[55] F. Shaffer and J. P. Ginsberg. "An overview of heart rate variability metrics and norms". en. In: *Front. Public Health* 5 (2017), p. 258.

[56] M. A. Hasnul et al. "Electrocardiogram-based emotion recognition systems and their applications in healthcare-A review". en. In: *Sensors (Basel)* 21.15 (2021), p. 5015.

[57] J. F. Thayer et al. "A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health". In: *Neurosci. Biobehav. Rev.* 36.2 (2012), pp. 747–756.

[58] C. Liu et al. "Dynamic Difficulty Adjustment in Computer Games Through Real-Time Anxiety-Based Affective Feedback". In: *International Journal of Human–Computer Interaction* 25.6 (2009), pp. 506–529.

[59] I. Homma and Y. Masaoka. "Breathing rhythms and emotions: Breathing and emotion". en. In: *Exp. Physiol.* 93.9 (2008), pp. 1011–1021.

[60] Q. Zhang et al. "Respiration-based emotion recognition with deep learning". In: *Comput. Ind.* 92-93 (2017), pp. 84–90.

[61] H. Liu et al. "Recent development of respiratory rate measurement technologies". en. In: *Physiol. Meas.* 40.7 (2019), 07TR01.

[62] C. Cortes and V. Vapnik. "Support-vector networks". en. In: *Mach. Learn.* 20.3 (1995), pp. 273–297.

[63] M. A. Cano Lengua and E. A. Papa Quiroz. "A systematic literature review on support vector machines applied to classification". In: *2020 IEEE Engineering International Research Conference (EIRCON)*. IEEE, 2020.

[64] N. H. Ovirianti, M. Zarlis, and H. Mawengkang. "Support vector machine using A classification algorithm". In: *SinkrOn* 7.3 (2022), pp. 2103–2107.

[65] R. Gholami and N. Fakhari. "Support vector machine: Principles, parameters, and applications". In: *Handbook of Neural Computation*. Elsevier, 2017, pp. 515–535.

[66] M. Schels et al. "Multi-Modal ClassifierFusion for the Recognition of Emotions". In: 10 (2013), pp. 73–97.

[67] C. J. C. Burges. "A tutorial on support vector machines for pattern recognition". In: *Data Mining and Knowledge Discovery* 2.2 (1998), pp. 121–167.

[68] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[69] M. Pal. "Multiclass approaches for support vector machine based land cover classification". In: (2008).

[70] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springe, 2006.

[71] M. Hejazi et al. "Multiclass support vector machines for classification of ECG data with missing values". en. In: *Appl. Artif. Intell.* 29.7 (2015), pp. 660–674.

[72] G.-X. Yuan, C.-H. Ho, and C.-J. Lin. "Recent advances of large-scale linear classification". In: *Proc. IEEE Inst. Electr. Electron. Eng.* 100.9 (2012), pp. 2584–2603.

[73] C. Hsu, C. Chang, and C. Lin. *A practical guide to support vector classification*. 2003.

[74] J. Shroff. *Player centric design*. https://www.thecreativemiddle.com/home/playercentricdesign/part1. Accessed: 2022-1-20. June 2018.

[75] R. Robinson et al. "Let's get physiological, physiological!: A systematic review of affective gaming". In: *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. New York, NY, USA: ACM, 2020.

[76] I. Kotsia, S. Zafeiriou, and S. Fotopoulos. "Affective Gaming: A Comprehensive Survey". In: *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2013.

[77] R. Lara-Cabrera and D. Camacho. "A taxonomy and state of the art revision on affective games". en. In: *Future Gener. Comput. Syst.* 92 (2019), pp. 516–525.

[78] W. Yang et al. "Physiological-based emotion detection and recognition in a video game context". In: *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018.

[79] P. A. Nogueira, R. Rodrigues, and E. Oliveira. "Real-time psychophysiological emotional state estimation in digital gameplay scenarios". In: *Engineering Applications of Neural Networks*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 243–252.

[80] S. Tognetti et al. "Enjoyment recognition from physiological data in a car racing game". In: *Proceedings of the 3rd international workshop on Affective interaction in natural environments - AFFINE '10*. New York, New York, USA: ACM Press, 2010.

[81] J. Frommel et al. "Integrated Questionnaires: Maintaining Presence in Game Environments for Self-Reported Data Acquisition". In: *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*. ACM, 2015, pp. 359–368.

[82] J. Frommel, C. Schrader, and M. Weber. "Towards emotion-based adaptive games: Emotion recognition via input and performance features". In: *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. New York, NY, USA: ACM, 2018.

[83] K. Gilleade, A. Dix, and J. Allanson. "Affective videogames and modes of affective gaming: assist me, challenge me, emote me". In: *Proceedings of DIGRA'2005*. 2005, pp. 1–7.

[84] M.-V. Aponte et al. "Scaling the Level of Difficulty in Single Player Video Games". In: *Entertainment Computing – ICEC 2009*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 24–35. ISBN: 978-3-642-04052-8.

[85] J. Chen. "Flow in games (and everything else)". en. In: *Commun. ACM* 50.4 (2007), pp. 31–34.

[86] G. Chanel and P. Lopes. "User Evaluation of Affective Dynamic Difficulty Adjustment Based on Physiological Deep Learning". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12196 LNAI (2020), pp. 3–23.

[87] M. M. Bradley and P. J. Lang. "Measuring emotion: The self-assessment manikin and the semantic differential"". In: *Journal of Behavior Therapy and Experimental Psychiatry* 25 (1994), pp. 49–59.

[88] M. Lombard et al. "Presence and television." In: *Human Communication Research* 26 (2000), pp. 75–98.

[89] J. Blascovich et al. "Immersive virtual environment technology as a methodological tool for social psychology". In: *Psychol. Inq.* 13.2 (2002), pp. 103–124.

[90] J. Diemer et al. "The impact of perception and presence on emotional reactions: a review of research in virtual reality". In: *In Front. Psychol* (2015).

[91] X. Peng et al. "A palette of deepened emotions: Exploring emotional challenge in virtual reality games". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2020.

[92] Y. Li, A. S. Elmaghraby, and E. M. Sokhadze. "Designing immersive affective environments with biofeedback". In: *2015 Computer Games: AI, Animation, Mobile, Multimedia, Educational and Serious Games (CGAMES)*. IEEE, 2015.

[93] O. I. Caldas, O. F. Aviles, and C. Rodriguez-Guerrero. "Effects of presence and challenge variations on emotional engagement in immersive virtual environments". en. In: *IEEE Trans. Neural Syst. Rehabil. Eng.* 28.5 (2020), pp. 1109–1116.

[94] L. Reidy et al. "Facial electromyography-based adaptive virtual reality gaming for cognitive training". In: *Proceedings of the 2020 International Conference on Multimodal Interaction*. New York, NY, USA: ACM, 2020.

[95] E. L. M. Naves et al. "Virtual and augmented reality environment for remote training of wheelchairs users: Social, mobile, and wearable technologies applied to rehabilitation". In: *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 2016.

[96] J. Gutiérrez-Maldonado, M. Rus-Calafell, and J. González-Conde. "Creation of a new set of dynamic virtual reality faces for the assessment and training of facial emotion recognition ability". en. In: *Virtual Real.* 18.1 (2014), pp. 61–71.

[97] O. Balan et al. "Sensors system methodology for artefacts identification in Virtual Reality games". In: *2019 International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*. IEEE, 2019.

[98] F. Abuhashish et al. "Emotion interaction with virtual reality using hybrid emotion classification technique toward brain signals". In: *International Journal of Computer Science Information Technology (IJCSIT)* 7 (2015).

[99] R. Somarathna, T. Bednarz, and G. Mohammadi. "Virtual reality for emotion elicitation – A review". In: *IEEE Trans. Affect. Comput.* (2022), pp. 1–21.

[100] E. Bekele et al. "Design of a virtual reality system for affect analysis in facial expressions (VR-SAAFE); Application to schizophrenia". In: *IEEE Trans. Neural Syst. Rehabil. Eng.* 25.6 (2017), pp. 739–749.

[101] J. Nam et al. "A new terrain in HCI: Emotion recognition interface using biometric data for an immersive VR experience". In: (2019). eprint: 1912.01177.

[102] B. J. Li et al. "A public database of immersive VR videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures". In: *Front. Psychol.* 8 (2017).

[103] I. Mavridou et al. "Towards an effective arousal detection system for virtual reality". In: *Proceedings of the Workshop on Human-Habitat for Health (H3) Human-Habitat Multimodal Interaction for Promoting Health and Well-Being in the Internet of Things Era - H3 '18*. New York, New York, USA: ACM Press, 2018.

[104] A. Felnhofer et al. "Is virtual reality emotionally arousing? Investigating five emotion inducing virtual park scenarios". In: *Int. J. Hum. Comput. Stud.* 82 (2015), pp. 48–56.

[105] P. Bilgin et al. "A comparative study of mental states in 2D and 3D virtual environments using EEG". In: *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019.

[106] G. Riva et al. "Affective interactions using virtual reality: the link between presence and emotions". en. In: *Cyberpsychol. Behav.* 10.1 (2007), pp. 45–56.

[107] C. Rojas et al. "Project us: A wearable for enhancing empathy". In: *Companion Publication of the 2020 ACM Designing Interactive Systems Conference*. New York, NY, USA: ACM, 2020.

[108] T. B. Alakus, M. Gonen, and I. Turkoglu. "Database for an emotion recognition system based on EEG signals and various computer games – GAMEEMO". en. In: *Biomed. Signal Process. Control* 60.101951 (2020), p. 101951.

[109] F. Pallavicini and A. Pepe. "Virtual reality games and the role of body involvement in enhancing positive emotions and decreasing anxiety: Within-subjects pilot study". en. In: *JMIR Serious Games* 8.2 (2020), e15635.

[110] K. Hidaka, H. Qin, and J. Kobayashi. "Preliminary test of affective virtual reality scenes with head mount display for emotion elicitation experiment". In: *2017 17th International Conference on Control, Automation and Systems (ICCAS)*. IEEE, 2017.

[111] C. Coelho et al. *Media presence and inner presence : The sense of presence in virtual reality technologies*. 2006.

[112] E. Brown and P. Cairns. "A grounded investigation of game immersion". In: *Extended abstracts of the 2004 conference on Human factors and computing systems - CHI '04*. New York, New York, USA: ACM Press, 2004.

[113] B. Meuleman and D. Rudrauf. "Induction and profiling of strong multi-componential emotions in virtual reality". In: *IEEE Trans. Affect. Comput.* 12.1 (2021), pp. 189–202.

[114] A. Kim et al. "Exploring the relative effects of body position and spatial cognition on presence when playing virtual reality games". en. In: *Int. J. Hum. Comput. Interact.* 36.18 (2020), pp. 1683–1698.

[115] E. J. Ciaccio. "Biomedical signal and image processing, second edition, review of biomedical signal and image processing, crc press, taylor & francis group, boca raton, review by edward j". In: *BioMedical Engineering OnLine* 12.1 (2013).

[116] D. Novak and other. "A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing". In: *Interacting with Computers* 24.3 (2012), pp. 154–172.

[117] K. Plarre et al. "Continuous inference of psychological stress from sensory measurements collected in the natural environment". In: *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks*. 2011, pp. 97–108.

[118] B. Chizi and O. Maimon. "Dimension reduction and feature selection". In: *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer US, 2009, pp. 83–100.

[119] D. H. Wolpert and W. G. Macready. "No free lunch theorems for optimization". In: *IEEE Trans. Evol. Comput.* 1.1 (1997), pp. 67–82.

[120] J. Marin-Morales et al. "Real vs. Immersive virtual emotional museum experience: A heart rate variability analysis during a free exploration task". In: *2020 11th Conference of the European Study Group on Cardiovascular Oscillations (ESGCO)*. IEEE, 2020.

[121] M. Granato et al. "An empirical study of players' emotions in VR racing games based on a dataset of physiological data". en. In: *Multimed. Tools Appl.* 79.45-46 (2020), pp. 33657–33686.

[122] L. B. Hinkle, K. K. Roudposhti, and V. Metsis. "Physiological measurement for emotion recognition in virtual reality". In: *2019 2nd International Conference on Data Intelligence and Security (ICDIS)*. IEEE, 2019.

[123] M. Moghimi, R. Stone, and P. Rotshtein. "Affective Recognition in Dynamic and Interactive Virtual Environments". In: *IEEE Trans. Affect. Comput.* 11.1 (2020), pp. 45–62.

[124] I. Shumailov and H. Gunes. "Computational analysis of valence and arousal in virtual reality gaming using lower arm electromyograms". In: *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017.

[125] S. Ishaque et al. "Physiological signal analysis and classification of stress from virtual reality video game". en. In: *Annu Int Conf IEEE Eng Med Biol Soc* 2020 (2020), pp. 867–870.

[126] J. Kim, E. Andre, and T. Vogt. "Towards user-independent classification of multimodal emotional signals". In: *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009.

[127] A. Anderson, T. Hsiao, and V. Metsis. "Classification of emotional arousal during multimedia exposure". In: *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments*. New York, NY, USA: ACM, 2017.

[128] J. A. Arnett and S. S. Labovitz. "Effect of physical layout in performance of the Trail Making Test". en. In: *Psychol. Assess.* 7.2 (1995), pp. 220–221.

[129] J. Schell. *The art of game design the art of game design: A book of lenses, third edition*. 3rd ed. London, England: CRC Press, 2019.

[130] I. G. Yildirim. "Time pressure as video game design element and basic need satisfaction". In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16*. New York, New York, USA: ACM Press, 2016.

[131] J. F. Hair. "Multivariate data analysis: An overview". In: *International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 904–907.

[132] A. Buja et al. "Statistical inference for exploratory data analysis and model diagnostics". en. In: *Philos. Trans. A Math. Phys. Eng. Sci.* 367.1906 (2009), pp. 4361–4383.

[133] J. W. Mauchly. "Significance test for sphericity of a normal n-variate distribution". In: *The Annals of Mathematical Statistics* 11 (1940), pp. 204–209.

[134] O. J. Dunn. "Multiple comparisons among means". In: *J. Am. Stat. Assoc.* 56.293 (1961), pp. 52–64.

[135] *VIVE Pro Eye overview*. en. https://www.vive.com/us/product/vive-pro-eye/overview/. Accessed: 2022-9-18.

[136] D. Micklewright et al. "Development and validity of the rating-of-fatigue scale". en. In: *Sports Med.* 47.11 (2017), pp. 2375–2393.

[137] B. Chandrakar, O. Yadav, and V. Chandra. "A survey of noise removal techniques for ecg signals". In: *International Journal of Advanced Research in Computer and Communication Engineering* 2 (Jan. 2013), pp. 1354–1357.

[138] C. Amiez and E. Procyk. "Midcingulate somatomotor and autonomic functions". en. In: *Handb. Clin. Neurol.* 166 (2019), pp. 53–71.

[139] M. Kelsey et al. "Artifact detection in electrodermal activity using sparse recovery". In: *Compressive Sensing VI: From Diverse Modalities to Big Data Analytics*. Ed. by F. Ahmad. SPIE, 2017.

[140] T. Van Steenkiste et al. "Systematic comparison of respiratory signals for the automated detection of sleep apnea". en. In: *Annu Int Conf IEEE Eng Med Biol Soc* 2018 (2018), pp. 449–452.

[141] D. Makowski et al. "NeuroKit2: A Python toolbox for neurophysiological signal processing". en. In: *Behav. Res. Methods* 53.4 (2021), pp. 1689–1696.

[142] A. Lourenço et al. "Outlier detection in non-intrusive ECG biometric system". In: *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 43–52.

[143] D. Khodadad et al. "Optimized breath detection algorithm in electrical impedance tomography". en. In: *Physiol. Meas.* 39.9 (2018), p. 094001.

[144] S. Raschka and V. Mirjalili. *Python Machine Learning*. en. 2nd ed. Birmingham, England: Packt Publishing, 2017.

[145] N. V. Chawla et al. "SMOTE: Synthetic minority over-sampling technique". In: (2011). eprint: 1106.1813.

[146] J. D. Rodriguez, A. Perez, and J. A. Lozano. "Sensitivity analysis of k-fold cross validation in prediction error estimation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.3 (2010), pp. 569–575.

[147] F. Krüger. "Activity, Context, and Plan Recognition with Computational Causal Behaviour Models". PhD thesis. Dec. 2016, pp. 70–73.

[148] T. N. Tombaugh. "Trail Making Test A and B: normative data stratified by age and education". en. In: *Arch. Clin. Neuropsychol.* 19.2 (2004), pp. 203–214.

[149] H. Song et al. "The effects of competition and competitiveness upon intrinsic motivation in exergames". en. In: *Comput. Human Behav.* 29.4 (2013), pp. 1702–1708.

[150] R. Bellman. "Dynamic programming". en. In: *Science* 153.3731 (1966), pp. 34–37.

[151] J. H. Brockmyer et al. "The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing". In: *Journal of Experimental Social Psychology* 45.4 (2009), pp. 624–634.