Investigation into Machine Learning and Emotional and Engagement Tracking Tools to Support and Enable At-Home Immersive Virtual Therapies

2024

Ryan Searle

University of Kent

School of Engineering

Page count: 210 (excluding bibliography)

Contents

List of Tables

List of Figures

Abbreviations

ML – machine learning

VR – virtual reality

PTSD – post-traumatic stress disorder

TRD – treatment-resistant depression

AR – Augmented Reality

rTMS - Repetitive transcranial magnetic stimulation

SVM – Support Vector Machines

KNN – K nearest neighbour

LR – Logistic Regression

DT – Decision Tree

RF – Random Forest

EHR – Electronic health records

IoT – Internet of things

VRET – Virtual reality exposure therapy

CBT – cognitive behavioural therapy

HCI – Human-computer interaction

CMA – circumplex model of affect

ANS – Autonomic nervous system

ECG – Electrocardiogram

HRV – Heart rate variability

GSR – Galvanic skin response

EMG – Electromyography

EEG – Electroencephalogram

CNN – Convolutional neural network

XAI – Explainable AI

LRP – Layer-wise Relevance Propagation

EFT – Emotion-focused therapy

HCI – Human-computer interaction

SAM – Self-assessment manakin

VAS – Visual analogue scale

P – Precision

R – Recall

EMG – Electromyography

ANS – Autonomic nervous system

SNS – Sympathetic nervous system

PNS – Parasympathetic nervous system

RAS – Reticular activating system

WHO – World Health Organisation

ET – Extra trees classifier

GB – Gradient boost

XAI – Explainable AI

CNN – Convolutional neural network

LSTM – Long short-term memory network

GEQ – Game engagement questionnaire

SHAP – Shapley additive explanations

LIME – Local interpretable model-agnostic explanations

Abstract

In recent years, the strain on the health system, exacerbated by the pandemic, has impacted mental health services, leading to growing pressures, shortages in mental health staff, and a lengthy waiting list for therapy. As of April 2023, there are 21,754 patients awaiting therapy, with 31.9% waiting for more than 18 weeks. To address these challenges, virtual therapy emerges as a promising solution, capable of alleviating stress on mental health services by offering frequent therapy and continuous monitoring. Utilising virtual reality (VR) technology, this form of psychotherapy creates simulated environments for treating various psychological conditions such as anxiety disorders, phobias, PTSD, and depression. Virtual therapy stands out due to its safe and controlled environment, high customisation potential and enhanced patient engagement. Virtual therapy has shown to be effective in treating many mental health disorders; advancements in technology, research, and accessible hardware have the possibility to expand on this research and create more personalised and remote therapies. To realise this future, ongoing investigations into mechanisms and technology for monitoring patients' reactions, feelings, and progress throughout therapy are essential. These mechanisms become even more important when you consider self-guided or automated at-home therapy.

This thesis presents several outcomes. First, we successfully constructed machine learning (ML) models capable of monitoring the mental health levels of patients diagnosed with treatment-resistant depression undergoing therapy, utilising data collected from Fitbit devices. Additionally, our next study demonstrated the effectiveness of VR in eliciting emotions compared to non-immersive stimuli. We established baseline ML results using a newly verified and published dataset, then enhanced these results through the implementation of deep learning techniques. To achieve real-time detection, we employed small data chunks. Addressing the interpretability challenge inherent in deep learning models, we developed a post hoc XAI system, offering visual explanations for predictions both locally and globally. These explanations were compared with medical and ML literature to gain deeper insights into the model's decision-making process. Furthermore, we investigated ML systems for detecting engagement during virtual experiences using EEG data, providing insights into feature importance and electrode placement. Overall, our research has not only investigated and initiated the development but also verified the viability of multiple tools essential for enabling immersive virtual therapies. Finally, we have offered guidelines on the utilisation of various physiological signals, hardware, and their most effective applications for future endeavours in this field.

Acknowledgements

I would like to express my sincere thanks to Dr. Chee Siang Ang (Jim) for giving me the opportunity to pursue this PhD and for his support and guidance throughout my studies. I am also grateful to my family and friends for their encouragement and support during this time; I would not have been able to complete this journey without them.

For Lou-Lou and Mark

Chapter 1: Introduction

1.1 Background

Recently, the healthcare system has faced an unprecedented strain, exacerbated further by the challenges wrought by the COVID-19 pandemic. The demand for mental health services, in particular, has surged, resulting in growing pressures, staff shortages, and an increasing backlog of patients awaiting therapeutic interventions. As of April 2023, 21,754 individuals found themselves on waiting lists for mental health services, with 31.9% waiting beyond 18 weeks (Baker, 2023). There is also an economic impact, in the UK, mental health problems cost at least £117.9 billion per year (McDaid et al., 2022). A significant portion of these costs arises from individuals with mental health conditions being unable to work or working reduced hours due to the extra challenges they encounter (McDaid et al., 2022). This situation underscores the need for innovative solutions in mental healthcare, paving the way for exploring new digital technologies

The strain on mental health services could benefit from adopting technological advancements. Mental healthcare services already use digital technologies to help deliver and improve therapeutic interventions, monitoring and diagnosis (Li, 2023). For example, cognitive behavioural therapy is being delivered online, resulting in greater accessibility and improved results (Ruiz-del-Solar et al., 2021). An example of anxiety diagnosis via digital technologies is presented by Sau and Bhakta, who used clinical data to predict anxiety in geriatric patients (Sau & Bhakta, 2017). This is an example of a system utilising the increasing amount of data produced and using artificial intelligence (AI) systems to provide insights.

Immersive and intelligent technology is something that is becoming more viable and could provide many additional benefits to the technology that has already been created. Immersive technology refers to technology such as virtual reality (VR) and augmented reality (AR) that "blurs the boundary between the physical and virtual worlds" creating a greater sense of immersion for users (Suh & Prophet, 2018). Throughout this thesis we will specifically be looking at VR. Intelligent technology refers to technology that utilises artificial intelligence (AI) or machine learning (ML). Immersive technology still needs to be fully realised; Future Care Capital reviewed the current technologies to aid mental healthcare, and less than 20% of the technologies utilised AI, and less than 10% used VR.

(Bloomfield & Ruiz de Villa, 2021). Immersive and intelligent technology combining VR and AI is promising, for example, training can be carried out for medical staff that simulates various scenarios they may face and utilise AI such as a speech recognition system to react to what the user is saying (Usmani et al., 2022). One area of specific promise and interest is in intervention and monitoring of mental health. One of the ways VR and AI can drastically improve mental healthcare is via immersive virtual therapy. Immersive virtual therapy, unlike traditional therapies, utilises VR to offer immersive, simulated environments tailored to address various mental health conditions such as anxiety disorders, phobias, post-traumatic stress disorder (PTSD), and depression (Geraets et al., 2021). This has multiple advantages: it provides a safe and controlled environment for patients, enables extensive customisation to cater to individual needs, fosters heightened patient engagement, and facilitates real-time feedback for therapists. While immersive virtual therapy has demonstrated success (Geraets et al., 2021), its full potential remains untapped, particularly in enabling at-home therapy.

At-home, immersive virtual therapy would allow patients to undertake adaptable and effective therapy at home without a clinician present. This could free up resources, reduce stress on clinicians and make therapy more accessible to people.

The future holds promise as advancements in technology and research will start to enable at-home immersive virtual therapy. As hardware becomes more affordable and accessible, at-home virtual therapy becomes increasingly possible. However, there are still research gaps, and investigation into the mechanisms and technology for monitoring patients' reactions, emotions, and progress throughout the therapy process is needed. This is because psychotherapy and the progression of treatments rely on monitoring these characteristics in patients throughout (Lutz et al., 2021; Sloan & Kring, 2007). Technology must be developed to track this and relay information back to the clinicians to enable at-home therapies where clinicians do not have to be present. To create these patient monitoring systems, signal data from physiological sensors needs to be leveraged. We could employ ML analysis to exploit this data fully.

Amidst the escalating challenges faced by mental health services, this research aids in exploring the development and integration of ML and emotional tracking tools for immersive virtual therapy. By exploring this domain, the thesis seeks to address gaps in tools required to enable immersive virtual therapy and, in turn, improve mental healthcare accessibility and effectiveness. After reviewing relevant literature in this field, we identified three gaps we want to investigate:

i) There is much research that looks into mental health tracking; however, most of these focus on healthy adults; there are limited studies that look at tracking patients diagnosed with mental health diseases throughout the course of therapy (see 2.8.1).

ii) Much research has looked into emotion recognition and, to a lesser extent, engagement recognition. However, these studies focus on using non-immersive stimuli; there needs to be more understanding of the efficacy of immersive experiences as emotional and engagement stimuli. As an extension to this, it needs to be better understood whether emotion and engagement can be reliably tracked during immersive experiences and how the classification models compare to those that classify emotion and engagement during non-immersive experiences (see 2.8.2).

iii) Further research must be carried out on XAI techniques and Interpretable ML models for emotion and engagement recognition. A specific need exists to understand models utilising raw time series physiological signals (see 2.8.3).

Therefore, in this thesis, we aim to investigate whether mental health can be tracked day to day throughout the course of psychotherapy, validate whether immersive experiences are reliable stimuli for emotion and engagement recognition and look at methods that can make emotion and engagement ML models interpretable for clinicians. This means providing interpretability on raw time series signals and providing context and validation of our models against medical literature.

The successful implementation of these technologies has the potential to begin to enable at-home immersive virtual therapy, offering timely interventions, personalised treatment approaches, and continuous emotional monitoring. Moreover, the study's findings promise to not only alleviate the strain on mental health services but also enhance patient outcomes and overall well-being.

1.2 Aims and Research Questions

This thesis aims to investigate the mechanisms and technologies that will begin to enable at-home immersive virtual therapy. To research and identify these, we needed to ascertain what is required for immersive virtual therapy to be successful. We need to verify that VR can elicit emotions required for therapy and further research physiological responses and predictive models that can recognise these emotions. For therapy to work without a clinician present, emotional state needs to be tracked throughout sessions and in day-to-day life throughout the course of treatment. ML analysis and physiological sensors will need to be leveraged to do this. Engagement in therapy also needs to be tracked to manage the therapy, get context to the results and provide meaningful feedback and therapy changes. It is essential that all predictions made by the AI models can provide reasons and explanations for them. Therefore, an XAI system needs to be built in a way that a clinician can use to understand why the model has made its prediction; this is required from an ethical and lawful standpoint, as future medical and treatment decisions could be made from the results of these models. This thesis will address this and the gaps in literature we identified by investigating the following research objectives:

1) *Can the mental health of patients diagnosed with treatment-resistant depression (TRD) be tracked daily throughout the course of therapy and monitored using off-the-shelf wearable sensors and ML techniques (Chapter 4)?*

In addressing this question, the study delves into wearable technology, exploring its potential as a tool for continuous mental health monitoring. By analysing data collected from patients diagnosed with treatment-resistant depression wearing off-the-shelf wearable sensors, the research aims to detect and monitor mental health states throughout therapy. The study seeks to establish and utilise wearable sensor data and ML techniques to recognise daily fluctuations in mental health, providing valuable insights into the feasibility of daily mental health monitoring outside traditional clinical settings. This aims to address gap i.

2) *Is VR a reliable stimulus for emotion (Chapter 5) and engagement (chapter 7) elicitation throughout virtual therapy?*

In investigating the efficacy of VR as a therapeutic medium, this question aims to better understand individuals' emotional and engagement responses during immersive virtual experiences. Through various VR experiences, these studies examine the emotional states evoked in participants and the levels of engagement they experience throughout. By analysing physiological responses and self-reported feedback, the research aims to validate VR as a potential tool for eliciting genuine emotions and engagement during therapy sessions. The findings contribute to our understanding of the immersive potential of VR in mental healthcare contexts. This aims to address gap ii.

3) *Can accurate ML models be developed to recognise engagement and emotions during immersive virtual experiences, and how do these compare with models that utilise non-immersive stimuli (Chapters 5 and 7)?*

This question delves into emotion and engagement recognition, a pivotal aspect of understanding patients' responses in virtual therapy settings. Chapters 4 and 6 explore the development of ML models capable of accurately recognising various emotions individuals express during VR experiences. By utilising various physiological signals, these studies endeavour to construct robust emotion recognition algorithms. The research addresses emotion and engagement recognition during immersive virtual experiences and how these models compare to those that utilise data from non-immersive stimuli. This aims to address gap ii.

4) *Can real-time emotions of VR users be detected using raw time series signals, and how can Explainable Artificial Intelligence (XAI) provide interpretable explanations for these predictions? (Chapter 6)?*

This question delves into the real-time aspect of emotion detection within VR environments. The study focuses on detecting and interpreting emotions throughout VR experiences. Furthermore, the research explores the integration of Explainable Artificial Intelligence (XAI) techniques to provide transparent and interpretable explanations for the predictions made by emotion detection models. Chapter 5 delves into the methodologies and algorithms used to achieve real-time emotion recognition and XAI techniques that can provide clinicians with context and information about these predictions. This aims to address gap iii.

5) *Can the physiological signals and their specific features that are indicators of emotion and engagement be identified? (Chapters 4, 5, 6 and 7)?*

Throughout this thesis, we continually look at the identification of physiological signals and their specific features indicative of emotions and engagement during virtual experiences, a question that is answered in Chapters 4, 5, 6, and 7. Throughout these chapters, our aim is not only to explore and pinpoint these physiological signals but also to provide insights and practical guidance on feature engineering, selection and importance. In addition to this, we also aim to compare these results to medical literature in order to provide more transparency and trust in the systems. This exploration seeks to benefit the future development of these systems, aiming to highlight the most relevant signals, features, and the underlying reasons for their importance. This aims to address gap iii.

1.3 Scope

This thesis investigates the technological infrastructure necessary to facilitate at-home immersive virtual therapies. The primary emphasis lies in examining technologies designed to monitor and track patients' emotional states and progress throughout the therapy duration and provide clinicians feedback.

Distinctly, the research does not centre on the development of immersive virtual therapy itself. Instead, the primary objective is to delve into the technologies supporting the effective implementation of at-home immersive virtual therapies. The thesis excludes testing these technologies during the immersive virtual therapy sessions.

In addition, the research aims to develop machine learning models, leveraging reliable stimuli, to enhance the robustness of emotion monitoring and progress tracking throughout the therapeutic process.

When developing ML systems for healthcare purposes, transparency and interpretability are paramount. Therefore, we will also explore how to make ML models more transparent and perform feature importance analysis to understand the underlying reasons for classifications better.

The scope encompasses exploring technologies that contribute to integrating at-home virtual therapies into mental health care practices, focusing on VR's ability to elicit emotions and engagement in users and the development of explainable machine learning models developed using reliable stimuli.

1.4 Contribution

The key contributions of this thesis can be summarised as follows:

1) *Demonstrating the efficacy of VR in Emotional Elicitation:* This thesis validates that immersive VR-based stimuli are more effective in eliciting emotions than non-immersive stimuli. This finding not only contributes to the understanding of emotional responses but also emphasises the potential of VR technology in psychotherapy.

2) *Establishing the Feasibility of Daily Mental Health Monitoring of Patients Diagnosed with TRD throughout the Course of Therapy with Off-the-Shelf Sensors:* This thesis showcases the

practicality of continuous mental health monitoring throughout the course of therapy using readily available sensors. The research paves the way for accessible and cost-effective solutions in mental health care.

3) *Identification of Machine Learning Models and Methods for Emotion and Engagement Recognition during immersive virtual experiences:* The research identifies and evaluates various machine learning models and techniques specifically tailored for emotion and engagement recognition. By providing an analysis of these methods, the thesis contributes to the advancement of emotion recognition technology, offering valuable insights for researchers and practitioners in the field.

4) *Demonstration of a Real-Time Detection Model and Explainable Artificial Intelligence (XAI) Framework that Utilises Raw Time Series Physiological Signals:* The thesis introduces a deep learning-based real-time detection model for accurately identifying emotions. Additionally, it proposes an XAI framework that detects these emotional states and provides interpretable visual results. This transparent approach enhances the credibility of emotion recognition systems, making them more trustworthy for clinicians and patients.

5) *Insights into Physiological Signals and Their Significance in Emotion and Engagement and a Comparison to Medical Literature:* The thesis investigates physiological signals and their underlying features and mechanisms. By exploring the importance of specific physiological signals in the context of emotion and engagement, the research provides valuable guidance for developing models that accurately predict these states. This insight enables more precise and reliable predictions and systems in the future. Transparency and trust in the ML systems developed will also be gained from this.

6) *Analysis of ML training and validation techniques for emotion and engagement recognition:* This research demonstrates that some methods can lead to overestimated performance metrics due to the failure to account for data leakage between the training, validation and test sets. By highlighting the limitations of certain approaches, the thesis advocates for more robust evaluation frameworks, such as subject-independent or leave-one-subject-out validation, to provide a more accurate and generalisable assessment of model performance.

Overall, this thesis contributes to the fields of emotion recognition, mental health monitoring, and human-computer interaction by offering novel findings, practical applications, and methodological insights. The combination of these contributions not only advances the academic understanding of detecting emotions and engagement but also has practical implications for the development of therapeutic technologies and mental health interventions.

1.5 List of Publications

The results obtained from this thesis have been published in peer-reviewed journals and conferences. A summary of these publications is provided in Table 1.1.

Table 1.1: List of Publications

Chapter	Journal/Conference	Title	Status
3	Journal of Affective Disorders Reports (2022)	Investigation of physical activity, sleep, and mental health recovery in treatment-resistant depression (TRD) patients receiving repetitive transcranial magnetic stimulation (rTMS) treatment	Published
4	UbiComp: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (2021)	VREED: Virtual Reality Emotion Recognition Dataset Using Eye Tracking & Physiological Measure	Published

1.6 Structure

The thesis is structured as follows:

Chapter 2: A review of the current literature is carried out. First, we look at the current problem overview and mental health statistics globally and within the UK. Digital technologies across healthcare are then looked at before looking more specifically at digital technologies within mental healthcare. The use of VR within healthcare is then explored. We then delve into the literature on emotion and engagement and the physiological responses to emotion and engagement. Following that we investigate ML within healthcare and mental healthcare before looking at emotion and engagement recognition. We finally then look at XAI and its uses in healthcare and emotion and engagement recognition before concluding with the gaps identified in the literature.

Chapter 3: This chapter outlines the methodology used to train and validate the algorithms and techniques proposed in this thesis. The data processing, training and validation procedures are explained. Attention is drawn to related literature that employs training and validation methods which may overestimate performance, underlining considerations that should be held when making comparisons, challenges specific to emotion and engagement recognition are highlighted. The chapter presents a summary of the datasets, modalities, and validation approaches employed throughout this thesis.

Chapter 4: In this chapter, we explore the feasibility of using Fitbit wearable sensors to monitor daily mental health accurately in individuals undergoing repetitive transcranial magnetic stimulation (rTMS) treatment for clinical depression. Our focus is on minimal features for continuous tracking during therapy, aiming for consistent and clinically meaningful results. The study involves twenty-four patients, and our objectives include analysing patient data, developing ML models to classify daily depression severity, and identifying key features crucial for accurate predictions.

Chapter 5: This chapter presents a study that involves the creation of a dataset wherein users' emotions are triggered through virtual environments delivered via a VR headset. The application of ML analysis is showcased to underscore the dataset's utility in emotion recognition research. The results are systematically compared with other widely recognised non-immersive emotion recognition datasets, providing insights into the effectiveness of the developed dataset and the virtual environments’ ability to evoke emotion. Additionally, the chapter delves into feature importance analysis conducted on the dataset's feature set, contributing information on the key aspects influencing emotion recognition outcomes.

Chapter 6: This chapter builds upon the previous research by focusing on developing a real-time emotion detection deep learning model. The foundation for this model is laid in the preceding chapter, which utilised data from the VREED emotion dataset, incorporating signals such as ECG, GSR, and eye tracking. The initial baseline ML analysis in the previous chapter yielded promising results. Chapter 5 seeks to enhance these findings by creating a deep learning model capable of real-time emotion detection, utilising short segments of data. Comparative analysis will be conducted between the results of this model and the baseline analysis. Additionally, the chapter introduces a methodology aimed at providing explanations and enhancing understanding of the model's predictions, employing both global and local explanations of the deep learning models.

Chapter 7: This chapter focuses on exploring and introducing an engagement detection ML model that leverages Electroencephalogram (EEG) signals. The EEG signals were captured during participants' engagement in tasks of varying intensity levels within a virtual environment. These results are then compared to similar studies that utilise non-immersive stimuli. The study looks into the examination of electrode significance and specific features in relation to engagement, providing insights into the key factors influencing the detection of engagement levels through EEG signals.

Chapter 8: This chapter serves as the culmination of the research journey, where findings, limitations, and future directions are examined and discussed. Each research question posed throughout the chapters is addressed, providing an overview of the study's outcomes. The discussion extends to the implications of the findings and how they contribute to the existing body of knowledge. Finally, the chapter outlines potential directions for future research.

Chapter 2: Literature Review

2.1 Overview

The prevalence of mental health disorders on a global scale is becoming pervasive, affecting millions of individuals across diverse cultures and societies. Ranging from anxiety to severe psychotic conditions, these disorders exert a significant influence not just on individual lives but also on the broader society as a whole. Their impact is multifaceted, extending beyond personal well-being to various aspects of society and economics.

As per the World Health Organisation (WHO) (WHO, Mental health 2023), depression stands among the leading causes of disability globally. Additionally, statistics reveal suicide as the fourth leading cause of death among individuals aged 15 to 29 years old. Those battling severe mental health conditions die prematurely, facing the potential of dying up to two decades earlier due to preventable physical health issues. WHO reports a trend of increasing mental health conditions and substance use disorders, marking a 13% rise in their prevalence over the last decade until 2017. Approximately 20% of the world's children and adolescents struggle with mental health conditions, emphasising the severity of the issue, with suicide ranking as the second leading cause of death within the 15-29 age bracket.

The pervasive impact of mental health disorders extends beyond individual health, presenting socioeconomic challenges that harm productivity, inflate healthcare costs, and diminish overall societal well-being. These conditions influence various facets of life, from impacting academic or professional performance to straining relationships with family and friends and hindering active participation in communities. The economic toll of just two prevalent mental health conditions, depression and anxiety, amounts to US$ 1 trillion each per year on the global economy. Despite these costs, the WHO reports that the global median of government health expenditure allocated to mental health remains below 2% (WHO, Mental health 2023), highlighting the disparity between the financial burden imposed by mental health conditions and the funding allocated to address them.

Annually, 1 in 4 individuals in England deal with some form of mental health problem (McManus et al., 2009). Additionally, within a week, 1 in 6 people reports experiencing common mental health problems like anxiety and depression (McManus et al., 2016). There has been a noticeable increase in the overall number of individuals reporting mental health problems; there has been a 20% rise in common mental health problems among both men and women from 1993 to 2014 (McManus et al., 2016). There is also an increase in severe mental health symptoms reported within a week, increasing from 7% to over 9% between 1993 and 2014 (McManus et al., 2016).

The landscape of mental health treatment in both England and Wales reveals some challenges: Reports from these regions indicate that only a third of adults dealing with common mental health problems receive any form of treatment, whether through talking therapies, medication, or a combination of both (McManus et al., 2016; Welsh Health Survey 2015). Notably, psychiatric medication emerges as the most prevalent treatment option, showcasing a reliance on pharmaceutical interventions (McManus et al., 2016). These issues result in increased pressure on clinical staff and staff shortages in mental health services, as Hillier et al. (2023) highlighted. There is also a high demand for mental health services, as outlined by Baker (2023), where there is a reported backlog of 21,754 patients awaiting treatment as of April 2023. Nearly a third of these individuals have been waiting for over 18 weeks, indicating delays in accessing mental health care.

Digital technology can support and form part of the solution to addressing these problems. Addressing the limitations in care capacity and quality, digital interventions offer a cost-effective solution (Gega et al., 2022). Digital health interventions hold promise in treating mental health conditions. Amid the COVID-19 pandemic, digital solutions provided a safer alternative to traditional face-to-face treatments (Philippe et al., 2022). With the demand for mental health services surpassing available resources, digital tools aid in identifying signs of distress and offering timely interventions. Moreover, these solutions offer increased accessibility for individuals in remote areas, those with limited mobility, and those preferring non-face-to-face interactions, ultimately making mental health services more inclusive and accessible (*Maximising the potential of digital in mental health.* 2023).

2.2 Digital Technologies in Healthcare

Digital health is a broad topic that encompasses a range of technologies (e.g. mobile health, telehealth, wearable devices (Center for Devices and Radiological Health, 2020)) that are used to deliver and support healthcare practices, improving various aspects of healthcare services (Sharma & Kshetri, 2020). The WHO's strategic vision states that digital health is a promising tool for ensuring universal access to high-quality healthcare services. Its implementation contributes to the efficiency and sustainability of health systems, enabling the delivery of affordable, high-quality, and equitable care (WHO, Digital Health 2023). The COVID-19 pandemic led to an increase in digital healthcare provision (Webster, 2020) and prompted a shift in consumer preferences and innovative care delivery methods (Shudes et al., 2023). Specifically in the UK, the Department of Health and Social Care and NHS England have prioritised the digital transformation of health and social care, recognising it as a top priority (GOV, A Plan for Digital Health and Social Care 2022).

The evolution of digital health, in which one of the earliest advancements was digitising medical records two decades ago, has transformed how healthcare is now delivered (Sharma & Kshetri, 2020). This shift spans technological advancements and extends to its accessibility and reach geographically (Chandwani et al., 2018). Its impact on healthcare delivery spans diagnosis, treatment, management, and prevention of health conditions (Sharma & Kshetri, 2020). Some of the more advanced systems now have rivalled human pathologists in diagnosing urothelial carcinoma (Zhang et al., 2019). Moreover, the strides made in artificial intelligence (AI), machine learning (ML), and deep learning have enabled tools capable of processing vast amounts of medical data and creating systems that can carry out complex actions, such as modelling biological systems to simulate diseases (Cuff, 2023). In our daily lives, the pervasiveness of digital health is evident with wearable technologies like smartwatches, for example, constantly tracking and analysing our health data.

2.2.1 Digital Technologies Across Healthcare

Digital health impacts and improves healthcare in many ways, from cost-saving measures to improved accessibility and patient care. Financially, it has proven to provide benefits; one such example is an internet-based treatment for adults with depressive symptoms, which shows a high likelihood of cost-effectiveness in treating depressive symptoms (Warmerdam et al., 2010). Moreover, its accessibility to diagnostic and medical services is increasing, as highlighted in various studies (Sayani et al., 2019; Sandberg et al., 2019). Implementing medical digital technologies has contributed to better healthcare accessibility, offering open information on health, treatment, and biomedical research progress. However, reliability, safety, testing, and ethical concerns persist (Senbekov et al., 2020). There are many benefits of digital technologies within healthcare, including improved patient care, enhanced overall health, and the potential to promote healthy behaviours through digital health interventions (Murray et al., 2016).

Technology plays a multifaceted role in enhancing healthcare across various domains:

i) *Administration and management:* the introduction of electronic health records (EHRs) in 1992 (Evans, 2016) has not only facilitated clinical research but has also enhanced patient care (Cowie et al., 2016) (see 2.5.1.4). In cardiovascular care, automated identification of at-risk patients, implementation of evidence-based guidelines for prevention, and increased connection between the healthcare provider and patient via accessible health records have all improved patient care (Roumia & Steinhubl, 2014).

ii) *Diagnostic and patient monitoring capabilities:* advancements have been made through technologies like ML, enabling real-time detection of conditions like breast cancer (Zhang et al., 2020) and image analysis systems that can detect oesophageal adenocarcinoma (Iwagami et al., 2020) (see 2.5.1.1). The advancements in healthcare mobile apps and wearable devices have also led to the constant monitoring of physiological parameters, aiding in disease diagnosis and treatment (Lu et al., 2020) (see 2.5.1.5). The increased ability to monitor has enhanced patient care by enabling early disease detection and intervention through digitally transmitted health data (Farias et al., 2020).

iii) *Surgical procedures:* robotics has had a significant impact on this area, especially in gynaecology, with FDA-approved robotic surgeries becoming commonplace (Gitas et al., 2022) (see 2.5.1.3).

iv) *Genomics and personal medicine* have seen improvements in accuracy and insights through technologies like DeepCpG (Angermueller et al., 2017).

v) *Training:* virtual reality (VR) and augmented reality (AR) have impacted healthcare training by providing immersive, controlled learning environments (Ricci et al., 2022) (see 2.3).

vi) *Large-scale analytics:* predictive analytics, fuelled by vast digitised clinical data, are shaping population health management by enabling a shift from reactive to proactive, personalised medicine (Guo & Chen, 2023), as seen in stroke prediction with 78% accuracy and a 19% miss rate (Dev et al., 2022) (see 2.2.2.4).

vii) *Drug design and discovery:* improvements have been made through the application of ML; one such example is a system that can generate novel molecules with similar physicochemical properties to known compounds (Segler et al., 2017) (see 2.5.1.2).

2.2.2 Digital Technology and Mental Health

The COVID-19 pandemic greatly increased society's use of technology devices for everyday tasks, work, education, and healthcare needs. This shift paved the way for digital therapeutics and evidence-based interventions to prevent, manage, or treat medical conditions (Weir, 2021). This technology could improve access to psychological help for individuals experiencing mental health issues globally (Bucci et al., 2019). Technological innovations should be considered as potential solutions to improve mental healthcare in the future. Digital platforms offer opportunities for self-monitoring and self-management, an alternative to traditional face-to-face or paper-based assessment methods and remove some of the constraints they possess (Bucci et al., 2019). The adoption of smartphones, now at 85% among UK adults, including those with severe mental health concerns (Firth et al., 2015), underscores the potential to bring treatment and services outside of a clinical setting, integrating them into individuals' daily lives irrespective of location or time constraints.

Digital technologies have emerged as valuable tools within mental healthcare, impacting diverse areas such as monitoring, intervention, management, diagnosis, and data analytics. The evolution of AI and ML has been making an impact, these are benefiting from the data collected from modern devices and the Internet of Things (IoT). IoT generally refers to a network of devices that can collect and share data (Sharma et al., 2018). This allows for the remote monitoring, control and automation of various processes. Wearable fitness trackers, smart phones and VR headsets have become more accessible and sophisticated, allowing these advancements. Below, we will present specific advancements and studies showcasing the utilisation of digital technologies in three aspects of mental healthcare: monitoring (2.2.2.1), intervention and management (2.2.2.2), diagnosis (2.2.2.3).

2.2.2.1 Monitoring

Various digital technologies are being researched and employed to help monitor mental health conditions. Some of the most common include technologies like remote video conferencing and instant messaging (Siegel et al., 2021; J. Devoe et al., 2022), which are commonly employed for support groups and appointments. Additionally, online programs, hotlines, and courses available through digital platforms (Murphy et al., 2021) contribute to mental health support. These are digitised versions of the traditional clinicians' monitoring of patients. These are forms of telemedicine, which is a technology that can provide remote healthcare support. Examples of telemedicine can range from using a telephone for consultations to more experimental technologies such as remote surgery (telesurgery) (Field, 1996).

With the global surge in smartphone ownership (99% of 16-24-year-olds in the UK) as of 2022 (Statista, 2023), remote sensing emerges as a practical approach for mental health management and treatment. Combining remote sensing with questionnaire completion offers more objective and frequent measures of mood and other crucial aspects of mental disorders, moving away from solely relying on patients' retrospective accounts (De Angel et al., 2022). Smartphones and wearable sensors offer continuous streams of data on various behaviours essential for mental health monitoring/assessment, such as:

i) Sociability: data such as average phone call frequency in a day and phone call duration have been shown to correlate with various mental health problems, such as stress (Cho et al., 2016).

ii) Sleep patterns: sleep features such as the number of awakenings and time spent in deep sleep can be monitored via wearables. These are discussed further in this chapter and section 2.5.2.

iii) Activity: activity data can encompass a range of features, including the number of steps and time spent sedentary. In its most basic form, motion has been used as a feature to predict depression (Ghandeharioun et al., 2017).

iv) Physiological: many physiological signals can be tracked and used to monitor mental health. These are explored more in section 2.5.2.

v) Location: GPS data such as circadian movement and location variance collected via mobile phones is strongly related to depression symptom severity (Saeb et al., 2015).

When considering specific data and features for collection, a review revealed that sleep features rank as the most utilised, followed by physical activity (De Angel et al., 2022). Associations were observed between lower sleep efficiency/quality and higher depression scores, emphasising the significance of sleep stability and time spent in bed, where longer durations correlated with elevated depression scores (De Angel et al., 2022). Moreover, higher depressive symptoms were linked to reduced engagement in physical activities (Lu et al., 2018). Features like location, sociability, and phone usage also emerged as relatively common aspects integrated into these tools (De Angel et al., 2022).

Validity and reliability (accuracy of the sensors) is also a concern around certain behaviours inferred from single sensors, like GPS for location (Adamakis, 2017) and accelerometer for movement and physical activity (Plasqui et al., 2013).

Large amounts of data can be used to analyse trends in populations as a whole and can provide valuable, wider-scale insights for mental health professionals. Utilising this data and predictive analytics has the potential to identify and predict mental health issues population-wide; some such issues that can be predicted are suicide, radicalisation and bullying (Beheshti et al., 2021). One of the largest sources of data able to aid in these systems is social media data (Kamran Ul haq et al., 2020). Research has shown that rules can be generated using unsupervised learning and large datasets to identify the presence of various mental health disorders, and the intern can then be used as a decision-support system within a clinical setting (Zhou et al., 2019). For more discussion on the use of ML in mental health monitoring, see 2.5.2.1.

2.2.2.2 Intervention and Management

When considering how intervention and management of mental health disorders are being impacted by digital technology, similar to monitoring, telemedicine is widely used, video conferencing and instant messaging are two of the most common uses. They are currently used for appointments and sessions (Chiesa et al., 2021; Drissi et al., 2021).

More experimental technologies and use cases, such as immersive technologies, are also being explored. Research into how VR can be used to improve therapy has been carried out since the late 1990s (Rothbaum et al., 1999; North et al., 1998). This type of therapy is called virtual immersive therapy and has been shown to be successful for various mental health disorders and phobias (Parsons & Rizzo, 2008; Persky, 2011; Lindner et al., 2020). A review on immersive VR therapy to treat Schizophrenia (Bisso et al., 2020) found that there were positive results when it is used to treat various psychotic symptoms, including delusions, hallucinations, or cognitive and social skills. Immersive VR therapy has also been compared to conventional therapy as a method to support recovery of depressive symptoms (see Figure 2.1) (Kiper et al., 2022), where the immersive VR therapy showed a significant reduction in depressive symptoms when compared to the conventional intervention. In their review, Bisso et al., (2020) identified five types of VR interventions:

i) *VR training:* This consists of VR environments that replicate everyday situations and challenges aimed to improve the cognitive abilities of individuals diagnosed with schizophrenia (La Paglia et al., 2016).

ii) *VR social skill training:* VR training aims to improve verbal, non-verbal, and cognitive skills. VR social skills training replicates traditional social skills training in a VR environment (Park et al., 2011).

iii) *VR avatar therapy:* This therapy involves a three-way conversation between a patient, clinician, and an avatar within the VR environment (du Sert et al., 2018).

iv) *VR CBT:* Uses the techniques of CBT and applies them to a virtual environment that can simulate and expose the patient to various scenarios (Pot-Kolder et al., 2018; Freeman et al., 2016).

v) *VR exposure:* Therapy that immerses the patient in increasingly scary or anxiety-inducing environments to desensitise them to their phobia/past experiences over time (Freeman et al., 2016).

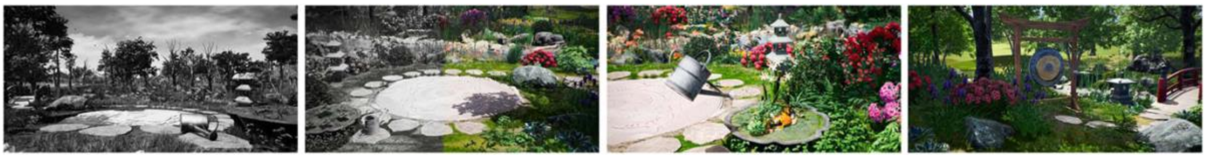


Figure 2.1: An example of therapy where a virtual environment has been used and gains colour over time, representing the patient's improvement) (Kiper et al., 2022)

Gamification is a topic that is currently being researched that can improve adherence and acceptability of various treatments and interventions. Gamification integrates game-like elements, such as competition, rewards, and interactive features, into non-game contexts or activities. The goal of gamification is to motivate behavioural changes and increase motivation for a task (Cheng et al., 2019). A commercial example of gamification to improve adherence to a task is Headspace (Headspace, 2024). Headspace is a meditation app that clearly shows the users' achievements and progress by displaying stats and using features such as 'badges' that can be earned by achievements, such as meditating for a certain number of days in a row. Headspace is one of the most downloaded well-being apps and has one of the highest retention rates; there are no studies that understand precisely how much the gamification features contribute to this, but it is sensible to believe they have some impact (Fleming et al., 2023).

Another technique used to engage and motivate users in digital health technologies is social features. There is limited research investigating the efficacy of social features in digital technologies; however, it has been shown that users can have a mixed response to social features; some feel that they feel motivated by certain competitive features, whereas others do not like the comparison to others (Tong & Laranjo, 2018).

Digital interventions overall are more accessible compared to conventional interventions that often require face-to-face meetings in physical locations; however, they do require a level of digital literacy, which is a barrier for practitioners and patients (Witteveen et al., 2022). It is also reported that in cases where the patient has a severe mental disorder, a reduced collaborative relationship between a patient and a therapist is felt (Witteveen et al., 2022). A review has also shown that digital forms of cognitive behavioural therapy (CBT) show comparable results to face-to-face psychotherapy; however, the acceptability of self-guided CBT is lower (Cuijpers et al., 2019). The previous statements on the effectiveness of CBT regarded non-immersive therapy, and the current literature on immersive virtual therapy is limited, but positive results are being published.

2.2.2.3 Diagnosis

When comparing digital technologies' impact on diagnosis to the previous two sections (monitor, intervention, and management), diagnosis is the least researched. There are a variety of methods to diagnose different mental health disorders; many are questionnaire-based. Many Current diagnosis methods have limitations, such as cultural and linguistic discrimination and subjectivity in interpretations (Roberts et al., 2018). Roberts et al., (2018) also suggest that digital technology could mitigate these limitations by providing three bits of data: self-reported, behavioural, and physiological. The ability to aid clinicians in making decisions via decision support systems (Tutun et al., 2022) using AI and questionnaire data is also currently being researched and showing promising results.

Mobile and other IoT technologies have not yet been implemented into routine diagnosis of mental health disorders (Roberts et al., 2018), but the current research is encouraging. Preliminary results of a diagnostic aid that uses data from dried blood spot proteomics to diagnose major depressive disorder are promising (Han et al., 2020). The research demonstrates the potential viability of these diagnostic aids in clinical practice. A questionnaire and AI-based decision support system has also achieved good results (Tutun et al., 2022). Using only 28 questions, the proposed decision support system could diagnose mental health disorders at 89% accuracy without human input. These evolving digital diagnostic methods showcase the potential to be implemented into routine mental healthcare and improve current diagnosis methods. ML is further advancing the classification and diagnosis of mental health conditions; see section 2.5.2.2 for more details.

2.3 VR for Healthcare

VR can help with mental and physical interventions and can be utilised for the training and education of clinical practitioners (Aziz, 2018). VR refers to a simulation that immerses users in a three-dimensional environment. These are most commonly simulated using a VR headset (see Figure 2.2), which allows users to look and sometimes interact with the environment. Early research focused on therapy for mental health conditions (Parsons & Rizzo, 2008; Persky, 2011), whereas today, more advancements are being made in surgical procedures (Wiederhold & Wiederhold, 2007; Aïm et al., 2016). Therapy has attracted a significant amount of research. Studies have shown VR-aided therapy to be beneficial for a range of impairments, including public speaking anxiety (Lindner et al., 2020), driving phobias (Trappey et al., 2020), improving upper limb function and the daily autonomy of stroke survivors (Rodríguez-Hernández et al., 2021).

Figure 2.2: Oculus Quest 2 VR Headset

VR can provide many benefits when implemented into healthcare settings; it can provide a more positive patient experience and allow clinical staff to make more informed decisions by, for example, allowing staff to practice procedures before carrying them out (Javaid & Haleem, 2020). Many factors contribute to improved patient experience, including the ability to customise and personalise the virtual environments, increased patient motivation and the accessibility of VR (Liu et al., 2022). A result of this is also an improved relationship between patient and therapist (Liu et al., 2022). There are also some limitations when looking at implementing VR into a clinical setting; a concern is that much of the research being produced is only directly relatable to research settings (Halbig et al., 2022), and adaptions need to be made before integration into practice.

Overall, VR has the potential to increase patient outcomes and healthcare delivery greatly. In terms of delivery, the implementation of VR into the training of clinical professionals has led to reduced errors and a reduced cost of training time and resources (Hanna et al., 2018). VR has also been shown to improve patient satisfaction in various areas of healthcare, including preoperative use (Bekelis et al., 2017) and therapy (Freeman et al., 2022).

As mentioned earlier, therapy is one of the leading fields of research within VR for healthcare. Treating anxiety disorders is one of the more advanced fields within VR therapy research (Geraets et al., 2021); however, there is much other research to see how VR can be used within therapy to treat psychotic disorders (Rus-Calafell et al., 2017), substance use disorders (Segawa et al., 2020), depression (Falconer et al., 2016) and eating disorders (Clus et al., 2018).

There are various ways VR therapy is carried out. VR exposure therapy (VRET) is a virtual version of traditional exposure therapy where patients are gradually exposed to their fears or traumatic experiences in a controlled environment; this has shown to be equally or more effective than other conventional forms of psychotherapies (Eshuis et al., 2021). Mindfulness and relaxation can also be carried out within VR (Yildirim & O'Grady, 2020). Patients can also carry out social skills training within VR (Zhou et al., 2021), where users can improve their social interactions, public speaking, etc, within a virtual environment. Pain management has also benefitted from VR; distractions can help manage chronic pain by exposing patients to immersive environments (Jones et al., 2016). These therapies are commonly therapist-guided; therapists can use VR as a tool during sessions, guiding patients through tailored experiences and providing real-time support and feedback.

One of the next steps in VR therapy is automated VR therapy that complements face-to-face sessions or acts as booster sessions after completing treatment; this could alleviate waitlists and make therapy more readily available (Geraets et al., 2021). Initial trials of stand-alone VR therapy are achieving promising results; one session of self-guided therapy aimed to improve public speaking anxiety has proven just as effective as one session of therapist-led therapy (Geraets et al., 2021).

There are some potential challenges to VR being a widely accepted tool within mental healthcare. Implementation into real-life practice is still limited (Brown et al., 2020). To make this a reality, it has been suggested that value in terms of cost-effectiveness, higher efficacy and treating patients unable to participate in conventional therapies needs to be demonstrated (Geraets et al., 2021). There is a positive outlook on the future implementation of VR into practice, as a study has shown that, in general, psychologists are positive about the use of VR (Lindner et al., 2019).

2.4 Emotion and Engagement for Mental Health Therapies

It is accepted that there are six basic emotions (anger, happiness, fear, surprise, disgust and sadness) (Ekman and Friesen, 1971). A three-dimensional model was developed in 1873, the three axes described are arousal, valence and intensity (Wundt, 1873). A two-dimensional model of this is used in most studies that are trying to classify emotion (Figure 2.3). This is known as the circumplex model of affect (CMA) (Posner et al., 2005).

In the context of psychology, arousal refers to the level of physiological and psychological activation or stimulation in an individual. Arousal can be thought of as a continuum, ranging from a low level of alertness or sleepiness to a high level of excitement or agitation. Arousal is “a state of physiological activation or cortical responsiveness, associated with sensory stimulation and activation of fibres from the reticular activating system." (*Apa Dictionary of Psychology* 2022).

In psychology, valence refers to the subjective positive or negative quality of an experience or emotion. Valence is a primary dimension of affect, or the experience of feeling, and can range from very positive to very negative. "in the field theory of Kurt Lewin, the subjective value of an event, object, person, or other entity in the life space of the individual. An entity that attracts the individual has a positive valence, whereas one that repels has a negative valence." (*Apa Dictionary of Psychology* 2022).

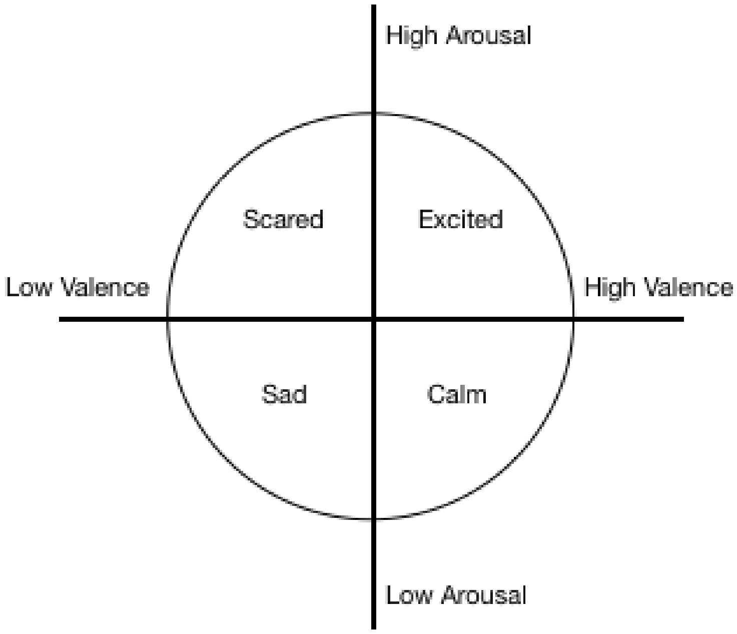


Figure 2.3: Circumplex Model of Affect (Posner et al., 2005).

Engagement is more complex to define than emotion. A broad definition is the level of involvement someone feels in a task (Ventura & Porfiri, 2020). Research has stated that engagement in humans consists of their activities, attitudes (Kappelman, 1995), goals and mental models, and motor skills (Said, 2004). Flow theory is strongly linked to engagement, which has been deemed a subset of flow in some research (Csikszentmihalyi, 1990). Flow is when someone is "so involved in an activity that nothing else seems to matter; the experience itself is so enjoyable that people will do it even at great cost, for the sheer sake of doing it" (Csikszentmihalyi, 1990, p. 4). A study that looks at a conceptual framework for defining user engagement with technology has also suggested that aesthetic theory, play theory and Information interaction all play a role in engagement (O'Brien &

Toms, 2008). Studies that classify engagement use various self-report questionnaires to define/measure the engagement a participant feels (Brockmyer et al., 2009; Nonis et al., 2020).

The monitoring of emotion throughout therapy is a factor in successful treatment outcomes. An example of how emotion is tracked and used in mental health therapy is during exposure therapy. Exposure therapy aims to expose patients to increasingly uncomfortable stimuli; systematically doing this allows the patient to gain a level of comfort in which was previously an uncomfortable situation (Öst et al., 1997). Therefore, comfort and emotion must be understood throughout the therapy to progress the stimulus systematically. Understanding emotions can also lead to improved outcomes for other mental health therapies, such as CBT (Samoilov & Goldfried, 2000) and emotion-focused therapy (EFT) (Greenberg, 2004). Research has shown that expanding the focus of CBT to include the understanding of the emotional meanings throughout can increase its long-term effectiveness (Samoilov & Goldfried, 2000).

Engagement also plays a major role in the effectiveness of therapies. The correlation between patient engagement in treatment and the psychotherapy outcome is well-established (Gaston, 1998; Gomes-Schwartz, 1978; Zuroff et al., 2016). Research indicates that treatment adherence may decline when patients are not actively engaged in their treatment, leading clinicians to deviate from treatment manuals during less engaged sessions (Snippe et al., 2018).

As we look at using digital technologies to enhance therapy, we must consider human-computer interaction (HCI) and how people interact and accept digitally delivered therapy. People treat computers the same way they treat people; therefore, computers should also respond humanely (Reeves and Nass, 1996). To do this, computers must understand their users' mental states and emotions. The ability to create applications that can respond and react to the users' emotions will allow applications to anticipate and address a patient's needs (Egger, Ley and Hanke, 2019). Using technology to predict human emotion can improve diagnosis and assessment of depression, prediction of treatment response, and early detection of response, remission and relapse (Ghandeharioun et al., 2017).

2.4.1 Physiological Responses to Emotion and Engagement

Human response to emotion involves a range of physiological and behavioural reactions. Behaviourally, emotions often prompt observable actions or expressions. For instance, a person feeling joy might smile. Physiological changes are more challenging for humans to control and consist of various underlying physiological responses such as heart rate and hormonal changes. Different states of emotion and levels of engagement elicit various physiological responses. Physiological patterns occur and are detectable (Rani et al., 2006) at different emotional and engagement states, even if the states are not expressed by the person by their gestures or facial expressions; this is called social masking. This happens because the sympathetic nerves of the autonomic nervous system (ANS) get activated when various levels of arousal or valence are experienced (Jerritta et al., 2011). The ANS is predominantly involuntary and is not easily consciously activated. Understanding and utilising these responses can lead to more objective emotional and engagement classification and monitoring.

Some of the most common physiological signals analysed in HCI and affective computing literature are (Jerritta et al., 2011):

- *Electrocardiogram (ECG):* the heart's electrical activity (see 4.2.3.3.2 for more detail).
- *Heart Rate Variability (HRV):* a measure of the variation in time between successive heartbeats (see 4.2.3.3.2 for more detail).
- *Galvanic Skin Response (GSR):* a measure of the skin's electrical conductance (see 4.2.3.3.3 for more detail).
- Respiratory System: features related to breathing, such as breaths per minute and respiration volume.
- *Electromyography (EMG): electrical activity of muscles.*
- *Electroencephalogram (EEG): electrical activity of the brain (see 6.2.1.2 for more detail).*

These can be used and are more powerful when used in combination. Multiple modalities can improve classification results in emotional recognition because two different modalities can be related to the two different axes on the CMA quadrant (arousal and valence). Sato et al., (2020) explored how EMG and GSR signals correlated to arousal and valence and found that EMG correlated with valence. In contrast, GSR correlated with arousal and complemented each other in a classification model. Multimodal data becomes even more significant when looking at deep learning, which performs well with large high-dimensional datasets which time-series physiological data can create.

As mentioned before, the integration of emotion recognition into CBT has been researched and is positive (Samoilov & Goldfried, 2000). Research is also looking into how physiological signals can inform therapy. Because physiological signals can give a more objective classification of emotion, clinicians can understand when patients are social masking or cannot report or provide insight into their emotions. Interpretable signals provide assurance and give the clinician confidence when situations such as social masking occurs (Field et al., 2015). These techniques can be multi-purpose and expanded into other therapies.

2.5 Machine Learning

The past decade has witnessed a growth in the use of ML techniques in broad aspects of human life. Healthcare is an area where applications of ML have grown significantly. With the adoption of EHRs in formal healthcare settings and the popularity of wearables such as Fitbits and Apple Watches in people's every day, a vast amount of data is being generated, which ML can take advantage of to provide improved care (Bhardwaj, Nambiar and Dutta, 2017). Here, we will provide an overview of ML and the various areas in which it is impacting the healthcare sector. Furthermore, we will discuss how ML is specifically being used within mental healthcare.

ML is a branch of AI that allows computers to learn from data and make predictions or decisions without explicit programming. Instead of manually coding every instruction, ML models identify patterns and relationships within large datasets. The process involves feeding data into algorithms that recognise these patterns, allowing them to generalise and make predictions on new, unseen data. ML is broadly categorised into supervised learning (learning from labelled data), unsupervised learning (finding patterns in unlabelled data), and reinforcement learning (learning from feedback or rewards).

This thesis focuses on supervised learning problems so here we present some of the most common techniques and the techniques used throughout the thesis:

K-Nearest Neighbours (KNN)

KNN is a non-parametric, instance-based learning algorithm used for both classification and regression tasks. It operates by comparing new data points to the labelled instances in the training set, classifying a new instance based on the majority label of its "K" nearest neighbours in the feature space (for classification) or averaging their values (for regression) (Murphy, 2012). As KNN does not make any assumptions about the underlying data distribution, it is highly flexible.

Decision Trees

Decision Trees are non-linear, tree-structured models that are applicable to both classification and regression tasks. The model works by recursively splitting the dataset into subsets based on the feature that maximises the separation between classes (or reduces error in regression) (Murphy, 2012). Each internal node represents a decision based on a feature, while leaf nodes represent the final outcome. A positive is that decision trees are easy to interpret and visualise.

Random Forest

Random Forest is a non-linear, ensemble method based on the aggregation of multiple decision trees. Each tree is constructed using a random subset of the data (bootstrapped samples) and a random subset of features (Biau & Scornet, 2016). Random Forest is more robust to overfitting when compared to decision trees and can improve performance by averaging the outputs (in regression) or using majority voting (in classification).

Support Vector Machine (SVM)

SVM is a supervised learning algorithm primarily used for classification, though it can also be extended to regression (SVR). SVM is a linear classifier that works by finding the optimal hyperplane that maximises the margin between different classes (Hastie et al., 2001). In cases where the data is not linearly separable, SVM can be extended to handle non-linear decision boundaries through the use of kernel functions, such as the radial basis function (RBF) or polynomial kernels. SVM is highly effective for high-dimensional spaces, but it can be computationally expensive for large datasets and requires careful tuning of the kernel parameters.

Logistic Regression

Logistic Regression is a parametric, linear classification algorithm that is widely used for binary classification problems. It models the probability of a data point belonging to a particular class using a logistic function (or sigmoid function) (Murphy, 2012). Logistic regression estimates the relationship between the input features and the log-odds of the outcome, assuming a linear relationship. While effective for linearly separable data, logistic regression struggles with more complex, non-linear relationships and requires the application of feature engineering or transformation techniques to handle such cases.

Gradient Boosting

Gradient Boosting is a powerful ensemble technique that builds models sequentially by iteratively correcting the errors of previous models (Hastie et al., 2001). Each model (usually a weak learner such as a shallow decision tree) is trained to predict the residual errors of the preceding model. This process leads to a strong predictive model by reducing both bias and variance. Gradient Boosting, as a non-linear model, can capture complex interactions between features. However, it is prone to overfitting if not properly regularised, and its training process can be computationally expensive, especially for large datasets.

Extra Trees

Extra Trees, is a non-linear ensemble method similar to Random Forest, but with greater randomness (Geurts et al., 2006). Unlike Random Forest, which selects the best split for each node, Extra Trees select splits by randomly choosing both the feature and the threshold for splitting.

Neural Networks and Deep Learning

Neural Networks are a class of non-linear models inspired by the structure of the human brain, consisting of layers of interconnected nodes (neurons) (Pal & Mitra, 1992). Each node in the network applies a mathematical function (an activation function) to its inputs and passes the result to the next layer. Neural networks are especially effective for capturing complex, non-linear patterns in large datasets. Deep learning, a subfield of neural networks, refers to networks with multiple hidden layers (deep architectures), allowing the model to learn hierarchical representations of the data. These models are particularly effective in domains such as image and speech recognition, though they require large amounts of data and computational resources for training. Additionally, due to their complexity, deep learning models are often considered "black boxes," making their interpretability a challenge. For more information on specific deep learning architectures such as convolutional neural network and long short-term memory networks, see sections 6.2.2.1 and 6.2.2.2.

2.5.1 Machine Learning in General Healthcare

Here, we will review various cases of ML being used in healthcare and look at advancements in diagnostics, drug discovery, robotic-assisted surgery, smart EHRs and sensors, and wearable technology.

2.5.1.1 Smart Electronic Health Records

EHRs produce large amounts of data for models to process and learn from. Most of this data is in the form of text and does not contain labels. Therefore, this data lends itself to natural language processing and unsupervised learning. Here, data is complex to model due to its noise, random errors and systematic bias (Weiskopf, Hripcsak, Swaminathan and Weng, 2013). However, unsupervised learning techniques such as auto encoding are starting to be used successfully to predict specific diagnoses (Miotto, Li, Kidd and Dudley, 2016). CNNs, recurrent neural networks (RNNs) (see 6.2.2.2 for more detail), and large language models (LMM) are also being used to predict incidents from the sequential data in health records, such as the order of events which had previously led to a specific incident (Choi et al. 2016). LMMs are ML models used for language generation and understanding. These can be made using RNNs (Zhang et al., 2023). Clinical voice assistants will likely be developed with the next generation of speech recognition (Shickel, Tighe, Bihorac and Rashidi, 2018). These will reduce the workload for clinicians and could provide more accurate data in EHRs. The main problem to overcome in developing this sort of system is the ability of a model to summarise the dialogue while understanding the important features of the conversation (Esteva et al., 2019).

2.5.1.2 Sensors and Personal Technology

ML's ability to learn over time provides an opportunity to provide more personalised care and monitoring for everyone through the use of sensors and wearable technologies such as smartwatches and phones. Three factors have allowed the hardware to reach a point where data collection and monitoring is reliable and affordable “1) increased data processing power, 2) faster wireless communications with higher bandwidth, and 3) improved design of microelectronics and sensor devices” (Andreu-Perez, Leff, Ip and Yang, 2015). These sensors initially focused on providing valuable insights for people, such as steps, heart rate, and exercise tracking. Now the technologies are also being used to tackle larger challenges such as diabetes management or elderly tracking (Kim, Campbell, de Ávila and Wang, 2019). One example of these technologies in use is Skinvision, a cancer assessment app on smartphone app that analyses images (de Carvalho et al., 2019). This app allows a user to take a picture of their skin lesion, which is then analysed by an algorithm. A risk rating is returned with a recommendation to see a doctor if the lesion is deemed’ high risk’. Validation of this app shows it to be successful, scoring 95.1% sensitivity in detecting premalignant conditions (93% for malignant melanoma and 97% for keratinocyte carcinomas and precursors) with a 78.3% specificity (Udrea et al., 2019). As the models and sensors improve, they will only become more accessible and beneficial for people to track and monitor their health. This will be particularly important in areas such as mental health, where there is a lack of resources (Lake, 2017)

2.5.2 ML in Mental Healthcare

2.5.2.1 Unobtrusive Mental Health Monitoring

Due to the advances in technology in mobile phones and smart wearable devices, researchers are starting to be able to detect and monitor mental health problems in more unobtrusive ways that do not interfere with the patients' daily life. Smart phones are being developed to have increasingly more sensors each year, which can be used for mental health detection. However, wearables, such as smart watches, can provide higher quality data as they can be worn all day and record throughout the day, unlike a phone that is often left and not used for large portions of the day. Accelerometers are the most widely used sensors in wearable devices, and they have been developed to track physical activities and exercise (Choudhury et al., 2008). According to a study based on 2,862 participants, accelerometer-based activities were strongly associated with decreased rates of depression (Vallance et al., 2011). This shows the value wearables could have to mental health researchers.

Sleep Disturbance is a well-known symptom of many mental health diseases (Sivertsen, Krokstad, Øverland and Mykletun, 2009), and accurate sleep detection is becoming a standard feature across many commercial wearable devices (Lee et al., 2018). Simple data, such as sleep duration, has been associated with depressive symptoms (Chen et al., 2013). Sleep states can now be detected as well, which correlate with severity of depressive symptoms (Wang et al., 2014). With the ability to track sleep more accurately and relatively cheaply, there is more research to be done.

With 66% of the UK population active on social media (UK: social media usage 2019 | Statista, 2020) and it being a place where users commonly share opinions and feelings, it could be beneficial as a modality to predict mental health issues. Information such as who a user is friends with, when they log on, and how many posts are made, can be used as well as linguistic analysis of speech (Rude, Gortner and Pennebaker, 2004) to detect different conditions. Depressed and non-depressed users can be determined by later posting times, less frequent posting, greater use of first-person pronouns, and greater disclosure about symptoms, treatment, and relationships on Twitter (De Choudhury, 2013). A sentiment analysis method using vocabulary and man-made rules calculates depression (Wang et al., 2013). A depression detection model was then made based on that method. Then, three kinds of classifiers were used to verify the model, and all achieved around 80% precision. In addition to this research, an application was developed using the proposed model for mental health monitoring online.

Similarly to social media, a decrease in phone calls, text messages and location entropy (a measure of the temporal dispersion of locations) were related to students feeling sad and stressed (Madan, Cebrian, Lazer and Pentland, 2010). The project Moodscope (LiKamWa, Liu, Lane and Zhong, 2013) used the number and length of communications (calls, SMS text messages, and emails), the number of apps used and usage patterns, web browser history, and a person's location to predict mood measured using ecological momentary assessments, with an accuracy of 66%. However, this accuracy improved to 93% after two months personal training period. This highlights the promising ability of machine learning to provide personalised care. Physiological sensors are also used to detect stress (Wijsman et al., 2011). ECG, respiration, skin conductance, and EMG were used. Nineteen features were extracted and normalised from these signals. Principal component analysis was then performed, where the features were reduced to 7 principal components. These features were then tested on various classifiers, classifying between stress and non-stress conditions with almost 80% accuracy, suggesting these features were promising for use in a personalised stress monitor.

2.5.2.2 Mental Health Classification and Diagnosis

Clinicians currently diagnose depression, and large parts of treatment rely upon the patient to self-report, monitor and self-administer medication which over half of patients fail to take adequate doses over an adequate period of time (Depression in adults - Monitoring | BMJ Best Practice, 2020). Technology could allow constant and closer monitoring of patients symptoms and medication adherence. Sensing has the potential to detect behaviours related to different mental health states and could start to uncover new information that could give researchers a better understanding of different mental health diseases (Mohr, Zhang and Schueller, 2017).

Many modalities are starting to yield promising results in the detection and monitoring of depression. Eye tracking has been proven to be able to detect depression with good accuracy. An affective sensing system to aid clinicians in their diagnosis and monitoring of depression been developed (Alghowinem et al., 2013). The system uses face videos to extract features from the eyes of patients to use in a binary classification task (depressed, non-depressed). 75% accuracy was obtained using statistical measures with support vector machine (SVM) over the full interview. Further investigation into the main differences between the depressed and non-depressed was carried out, and it was found that there was no difference in blink rate between the two classes. However, the average distance between eyelids was significantly smaller, and the average duration of blinks was longer in depressed patients, suggesting fatigue or eye contact avoidance.

Eye tracking and facial tracking are closely linked in method and use. Unsurprisingly, facial expressions have also been studied in their relation to depression. One study aimed to classify and diagnose depressed patients using ML methods to analyse videos (Wang, Yang and Yu, 2018). Twenty-six healthy participants and 26 hospitalised patients diagnosed with depression were filmed reacting to images. Certain points in the eyebrows and corners of the mouth were tracked along with pupil movement and blink frequency, and statistical features were derived from the videos. The depressed patients were classified using an SVM model with 78.85% accuracy.

Facial expression and movement have been combined with audio analysis for automatic depression analysis (Joshi et al., 2013). This study highlighted the importance of data fusion. This paper looked into three different fusion types: feature (concatenated features from multiple modalities), score (combined scores such as probabilities and likelihoods are combined and make a classification decision) and decision fusion (multiple classifiers trained on different feature sets). The experiment consisted of healthy and depressed patients watching movie clips, watching and rating international affective picture systems, reading sentences containing affective content and an interview. The audio features extracted were from the secondary channel and performed better than the video features. If the primary channel was analysed (what words were said), the accuracy could be improved further. The three different fusion types outperformed the singular modality models and performed at similar levels to each other.

All methods mentioned above are based on a singular test or interview. The technology discussed following makes use of wearable sensors and can be measured throughout the day in an unobtrusive way. One of the most basic in terms of technology needed is the use of GPS and phone usage data from participants smartphones to identify participants with depressive symptoms with 85.6% accuracy (Saeb et al., 2015). This study found that the number of locations a person visited did not relate to depression, but the variability in time spent in different locations did. This means the more time a person spent in few locations, the more likely a person was depressed.

Physiological signals are also being used to detect depression. A specific hardware system has been developed to classify depressed patients (Roh, Sunjoo Hong and Hoi-Jun Yoo, 2014). Using this sensor to read an ECG signal, time-domain frequency-domain and non-linear HRV features are extracted. This system can perform most of the complicated signal analysis functions within the hardware fast and with low power consumption. These features are then sent to a mobile app where an SVM classifier achieves a classification accuracy of 71%.

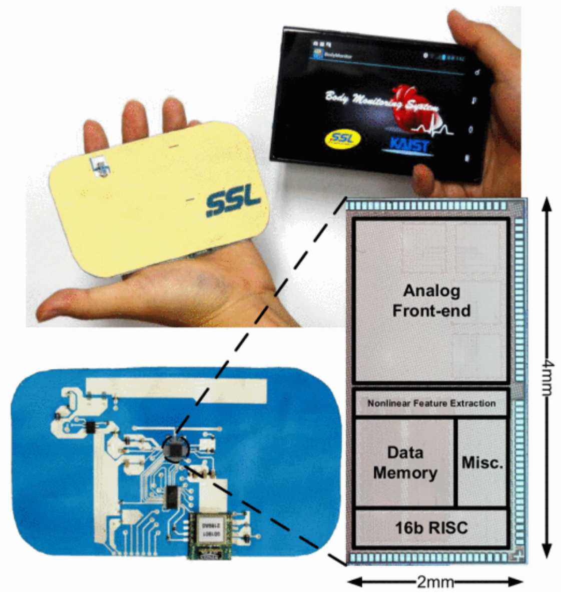


Figure 2.4: Hardware proposed by Roh Sunjoo Hong and Hoi-Jun Yoo (2014) to monitor depression.

EEG systems have also been developed to help monitor a patient's electrophysiological signals more conveniently for a doctor. Certain EEG features are shown to be heavily related to depression, and correlation dimension, among other nonlinear features, has achieved a classification accuracy of 83.3% (Hosseinifard, Moradi and Rostami, 2013). A real-time system that could has proven to diagnose depression accurately (Zhao et al., 2017). This system could speed up the time it takes to diagnose a patient and the access that patient gets to the correct treatment by allowing users to self-test at home.

Rapid diagnosis has also been a subject of research. The process normally consists of questionnaires that could last days or weeks. A fear induction system for children has been proposed that lasts 90 seconds and achieves a diagnostic accuracy of 80% (McGinnis et al., 2018). This study measured motion while being exposed to something scary, a snake in this particular experiment. With the results presented and the significant reduction in time and cost compared to current methods, this study provides one possibility for future of depression diagnosis.

2.5.3 Challenges of Implementing ML in Healthcare

There are many benefits and use cases of ML in healthcare; however, we must also consider the limitations and problems faced when developing and utilising these systems. A large amount of data is being created; however, data quality must be considered. The data collected might lack representation from the populations for which the model is intended to be trained. For example, there are many studies that have researched monitoring depression levels; however, most of these studies focus on students or healthy adults. The insights and models derived from such data may not be directly applicable to patients diagnosed with depression, the demographic that a real-world depression monitoring system could target.

With the increase in data, deep learning models are starting to become the most affective and widely used models (Esteva et al., 2019); however, these suffer from a lack of transparency. These models are impossible for humans to interpret and can cause problems in certain situations if the reason for a decision cannot be explained. This problem is common across many other ML models as well. Without this understanding of inner workings, there can also be a lack of trust in these algorithms in making the correct decision or diagnoses and in how they will handle private data (Luxton, Anderson and Anderson, 2016). In addition, there are also times when a patient may receive sensitive information or an outcome from an AI system that would be better served by an empathetic healthcare worker (Davenport and Kalakota, 2019).

There is also no clear guidance on who will take accountability if the AI makes a mistake (Davenport and Kalakota, 2019). In a problem not unique to healthcare, algorithms can suffer from an inherent bias learnt from data. One case of this outside of the healthcare domain is from an algorithm developed to aid judges sentencing, which showed a disturbing tendency towards racial discrimination (Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say, 2020). These biases could also be unknowingly built into healthcare algorithms (Char, Shah and Magnus, 2018). Even with these problems, major improvements in healthcare are being made due to these models and increasingly more are being implemented with great success in many domains.

2.6 Emotion and Engagement Recognition

2.6.1 Emotion Recognition

There are multiple ways for humans and machines to identify emotion, one of which is speech. Speech consists of two channels, primary and secondary. The primary is what we say, whereas the secondary is the paralinguistic information, such as the tone and emotional state (Casale, Russo, Scebba and Serrano, 2008). Commercial emotion recognition models that utilise speech are available. With deep learning, five different emotions can be classified with an accuracy of 83% (Vokaturi, 2023). A more basic and less computationally expensive approach using the Bayesian classifier achieved 57.1% accuracy when classifying eight different emotions, with particularly high accuracy when classifying anger and sadness (Castellano, Kessous and Caridakis, 2008). For speech to be successfully implemented into real-world situations and improve, these systems will need to be able to work in noisy environments and start to recognise words and more complex aspects of speech such as sarcasm (Garcia-Garcia, Penichet and Lozano, 2017)

Computer vision techniques can analyse facial expressions from images or videos. Microsoft has released a commercial API (Facial Recognition | Microsoft Azure, 2020) that will Analyse images or videos and return the presence of each of Paul Ekman's six basic emotions (Ekman, 1992) on a scale from 0 to 1. However, there is a lower degree of accuracy and a significant drop in the actual detection of a face when not looking straight at the camera (Khanal et al., 2018). Emotion recognition systems are often built on artificial datasets with faces and obvious expressions. Research has been done into facial recognition with 3d cameras to recognise emotion even when it is not clear (Zhang, Cui, Liu and Zhu, 2016). This research proposes a new method to identify three kinds of emotion (happy, sad, neutral). A Kinect device is used to capture 1347 facial points before feature selection is conducted. This system achieved an accuracy of 70%, 77% and 80%, respectively for sad, happy and neutral. Similar problems arise when detecting emotion through facial expressions and speech; background noise can interfere. Even though Zhang et al. (2016) achieved good results, interferences from the environment, such as illumination, humidity and temperature, were not considered, which could hinder its usefulness in a real-world application.

One domain proven to acquire higher accuracy than the previous two is body gesture and movement (Castellano, Kessous and Caridakis, 2008). Even though we do not actively use our body gestures to communicate, they are used similarly to the second channel in our voices (Body Language: Picking Up and Understanding Nonverbal Signals, 2020). Castellano et al. (2008) extracted five motion cues using EyesWeb Expressive Gesture Processing Library (Camurri, Mazzarino and Volpe, 2004). A Bayesian classifier was then used to classify the emotions. This classifier achieved 67.1% accuracy. Anger and pride were recognised with the highest accuracy (80 and 96.67% respectively). In contrast, sadness was more challenging to recognise (36.67%). This study used a restricted set of data from a small group of subjects, which raised questions about how it would generalise, but it still showed promising signs. Combining all features gained by far the best performance with a 78.3% accuracy. Combining features is common across literature and should guide future works.

With social media and text messages becoming ever more ubiquitous, it is becoming more important and useful to recognise emotion through text. There are generally two approaches to this problem: feature-engineered-based approach (Balahur, Hermida, and Montoyo, 2012) and deep learning-based approach (Abdul-Mageed and Ungar, 2017). Deep learning methods are much more common now and achieve better results (Chatterjee et al., 2019). Chatterjee et al. (2019) achieved a model that significantly outperforms traditional ML baselines and other deep learning models using semantic and sentiment-based representations. Using text-based classification may be accurate but cannot be used for real-time detection, for this, we need other modalities, one of which being physiological signals.

Physiological signals such as ECG, EMG, EEG, GSR, blood volume pressure, heart rate or HRV, temperature and respiration rate are used to detect emotions (Agrafioti, Hatzinakos and Anderson, 2012). “The amygdala generates emotional impulses which create the physiological reactions associated with emotions” (Emotions and Physiology | alive, 2020). EEG and heart rate measurements can recognise valence and arousal, and respiration rate can be used to recognise certain states such as panic, fear, concentration, or depression. For this reason, it is recommended that these modalities be combined to detect the full spectrum of emotions (Egger, Ley and Hanke, 2019). With smart devices and wearables becoming widely used and unobtrusive, physiological signals are easier to measure on the go, allowing for emotion recognition during everyday activities (Egger, Ley and Hanke, 2019). In 2008, an emotion recognition system for race car drivers was developed using EMG, ECG, RR and electrodermal activity that achieved a 79.3% accuracy for five emotional states (high stress, low stress, disappointment, euphoria, neutral). This system highlights how accurate emotion detection can be developed for situations where the subject is in motion (Katsis, Katertsidis, Ganiatsas and Fotiadis, 2008).

2.6.2 Engagement Recognition

Engagement recognition has received less attention than emotional recognition; however, some studies demonstrate this task's feasibility. The investigation of engagement has been predominantly conducted through statistical and ML analyses, with video games as a primary instrument for data collection and examination in this context (Gábana Arellano et al., 2016; Chanel et al., 2011). Various methodologies have been employed in studies to elicit engagement, with some particularly successful approaches centring on manipulating competitiveness levels within gameplay scenarios. For instance, participants have been subjected to playing the same game individually, collaboratively, and competitively, demonstrating the versatility of this experimental paradigm

(Gábana Arellano et al., 2016). Additionally, manipulating game difficulty has proven effective in inducing engagement, with the rationale being that excessive ease leads to boredom, excessive difficulty induces anxiety, and an optimal level of difficulty fosters engagement (Chanel et al., 2011; Chanel et al., 2008).

Engagement detection has leveraged various physiological signals, with EEG signals emerging as a commonly employed modality. Rogers et al., (2020) employed a single left pre-frontal electrode to measure engagement in virtual reality therapy, finding that frontal theta power in healthy adults is a valid metric for user engagement within VR. Monteiro et al. (2018) investigated engagement in VR games by comparing first and third-person perspectives, determining that while there was no discernible pattern in signals based on viewing perspective, the signals offered insights into user preferences and perceived difficulty levels within the games. Another study by Abadi et al. (2013) utilised frontal EEG for engagement classification in affective cinema, exploring the distinct contributions of GSR, EEG, and facial tracking. The findings revealed that each modality significantly encodes participant engagement across a spectrum of video clips. The practical consideration of utilising EEG signals in VR-based experiences is underscored by the potential integration of EEG sensors into VR headsets, as these devices are already being developed and sold primarily for research purposes (Ag, 2023).

2.7 XAI for Emotion and Engagement Tracking within Healthcare Systems

The development of ML models, especially deep models, has led to increasing performance over time but has also led to a greater degree of complexity and a decrease in transparency regarding why a prediction was made. These hard-to-interpret models are referred to as “black boxes”. From this problem, explainable AI (XAI) has emerged as an important topic and area of research. XAI aims to explain models to humans, improving the trust and transparency of AI systems (Gerlings et al., 2021). As AI progresses, it will be used to help make more critical decisions, so stakeholders are increasingly demanding the transparency of models (Preece et al., 2018). An example of where explanations are key is in healthcare, where AI is used to support diagnosis (Tjoa & Guan, 2021).

In the existing literature on XAI, various criteria are employed to categorise XAI approaches (Theissler et al., 2022): ante-hoc and post-hoc, global and local, model agnostic and model specific. Ante-hoc methods consist of models that can be directly interpreted due to the design such as decision trees. Post-hoc approaches are separate from the model and do not change the model's underlying structure. Examples of post-hoc methods include Shapley additive explanations (SHAP) and local interpretable model-agnostic explanations (LIME) (see 6.2.3). SHAP is an XAI method based on game theory that measures the marginal contributions of each feature to a particular prediction. SHAP values are computed by evaluating the contributions of individual features within the context of all possible feature combinations in a prediction. This involves assessing the incremental impact of each feature when introduced into a group of other features. The resulting Shapley value ensures an equitable attribution to each feature, accounting for its interactions with other features. LIME is another XAI method that obtains its results by generating perturbed samples around a particular data point, obtaining predictions from the black-box model for these samples, and then fitting a locally interpretable model (such as a linear regression model) to approximate how the black-box model behaves in the vicinity of the chosen instance.

Ante-hoc methods are not suitable for most of the work presented in this thesis because the research employs multiple ML algorithms for different tasks, necessitating the use of model-agnostic explainability techniques. Ante-hoc methods integrate interpretability directly into the model architecture itself, making them specific to certain models and limiting their flexibility across a diverse range of algorithms. For instance, Decision Trees and Linear Models are inherently interpretable and serve as examples of ante-hoc methods, providing clear insight into how input features influence predictions. In Decision Trees, each internal node represents a feature-based decision rule, and each leaf node represents an outcome, allowing users to easily trace how input features lead to specific predictions. Generalised Additive Models (GAMs) (Hastie and Tibshirani, 1990) are another well-known ante-hoc technique, which combines interpretability with some flexibility by allowing non-linear relationships between features and outputs while maintaining transparency.

Global and local explanations refer to what is being explained. Global methods return an explanation for the whole model and all of the data points, giving a general explanation of the inner workings of a model. Local explanations explain a single prediction/instance. For instance, in the case of a model predicting a specific disease, global explanations offer insights into the overall population, highlighting factors influencing the prediction of that disease for the entire group. Whereas a local explanation provides context and reasoning for predicting the disease for an individual patient.

Agnostic methods refer to post-hoc methods and whether they have been developed specifically for a model or can be used across all models. For example, a model-agnostic method such as SHAP (Lundberg and Lee, 2017) could be used to explain predictions given by a random forest (RF) or an

SVM, etc whereas a model-specific method such as gradient class activation maps (grad-CAM) (Selvaraju et al., 2017) can only be used to interpret differentiable classifiers like CNNs. Grad-CAMs are a local XAI method to visualise and understand deep learning models. They result in a class activation map that highlights areas of the input data that were important to a classification (see Figure 2.5). See table 2.1 for a comparison of these XAI methods.

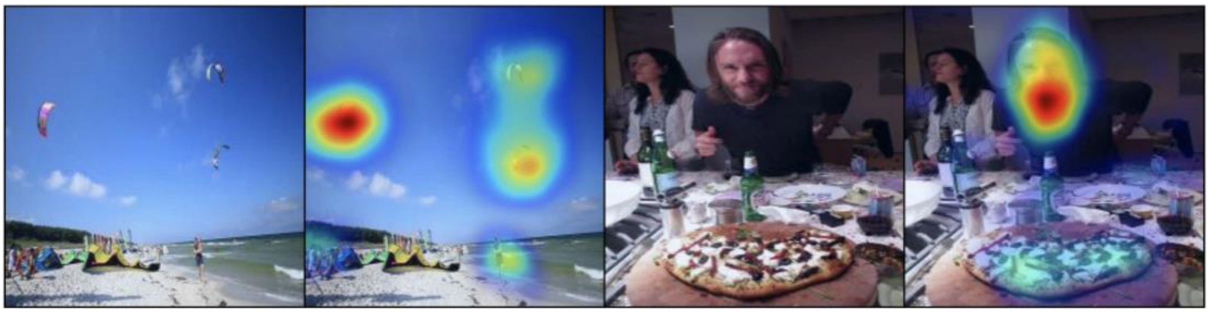


Figure 2.5: Example of a class activation map (Selvaraju et al., 2017)

Table 2.1: XAI Method Comparison

XAI Method	Pros	Cons
SHAP	• Model agnostic • Provides local and global explanations	• Computationally expensive
LIME	• Model agnostic • Computationally efficient	• Cannot provide global explanations
Grad-CAM	• Provides intuitive visual explanations • Computationally efficient	• Limited to deep learning models • Cannot provide global explanations

2.7.1 XAI in Healthcare

As mentioned in the prior section, using AI to aid critical decisions in healthcare has led to increased demand for XAI. Recent research indicates that for the successful integration and adoption of AI models in healthcare practices, it is important for models to be transparent and interpretable (Cai et al., 2019; Keane & Kenny, 2019). Research has suggested that XAI would increase stakeholders' willingness to utilise AI within healthcare due to the increased trust and understanding XAI can provide (Ribeiro et al., 2016; Holzinger et al., 2017; Pawar, 2020).

XAI is currently being researched and used within the healthcare domain. A systematic review of the last decade of XAI within healthcare highlighted that XAI is being used not only to explain results but also to verify model performance (Ribeiro et al., 2016) and guide the hyperparameter tuning (Loh et al., 2022). Some common techniques stated were LIME and SHAP for clinical feature explanation and grad-CAM for medical imaging. Loh et al., (2022) highlighted a relevant statement: there is a lack of research on 1D bio signals abnormality explanations. Research findings have highlighted advantages associated with XAI, including its potential to improve decision confidence among clinicians and contribute to the formulation of hypotheses regarding causality. This, in turn, contributes to heightened reliability and acceptance of the system within the healthcare domain (Antoniadi et al., 2021).

As mentioned previously, LIME, SHAP and class activation maps (CAM) are three of the most common XAI techniques used within the healthcare domain (Loh et al., 2022), here, we will present a couple of studies that employ these techniques. SHAP has been used in research to understand the importance of features to a model to predict cardiac surgery-associated acute kidney injury (see Figure 2.6) (Tseng et al., 2020). Cardiac surgery-associated acute kidney injury was successfully predicted, and visualisations created utilising SHAP highlighted the positive and negative effects of the top twenty features and were also used to explain how a single feature affected the model. The information gained from the XAI results was compared to medical literature to give context and confidence to the model. CNN models that detect metases have been explained using LIME (see Figure 2.7) (Palatnik de Sousa et al., 2019). LIME was used on multiple publicly available CNN models, and the explanations were compared to medical literature; it was found that the models did use some features a clinician would use to classify lymph node metases. CAMs have been used to identify sections of ECG signals relevant to predictions (see Figure 2.8) (Goodfellow et al., 2018). This research aimed to classify different ECG rhythms for a given waveform and then used CAMs to explain the predictions. The visualisations highlighted areas of the waveform that the model focused on when making a prediction. The highlighted areas fit the clinical understanding of the various rhythms and could provide a level of interpretability for clinicians working in a clinical context. This is not an exhaustive list of methods or use cases, but it provides context as to how some of the more popular methods are being utilised currently.

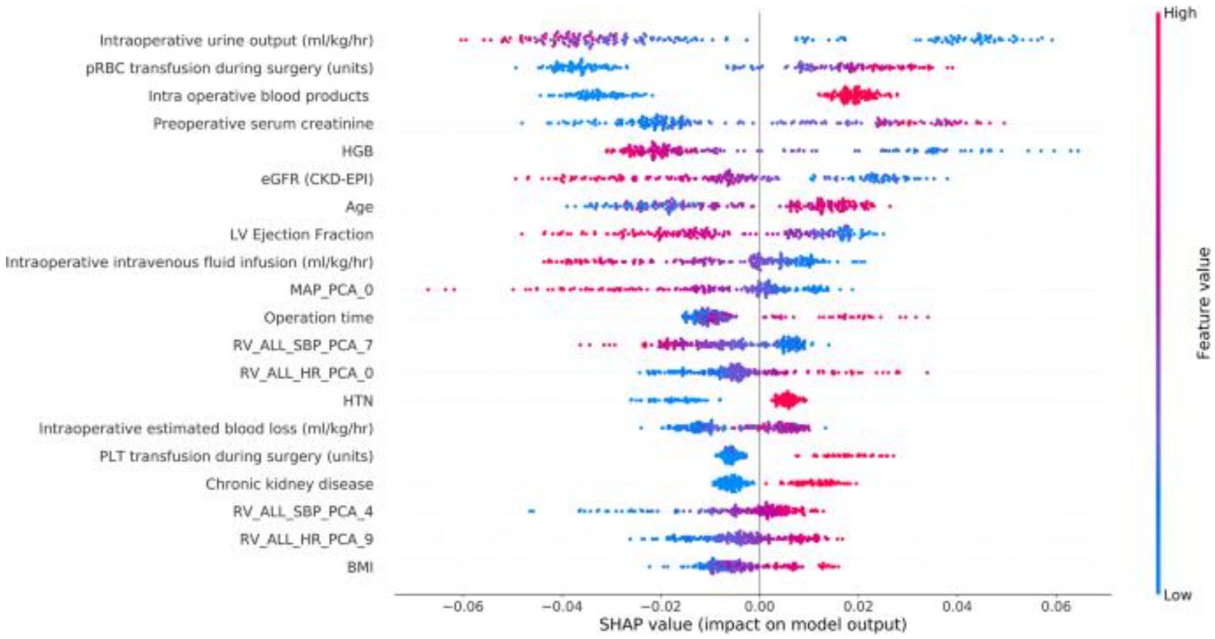


Figure 2.6: SHAP plot that shows the importance of features when predicting cardiac surgery-associated acute kidney injury (Tseng et al., 2020)

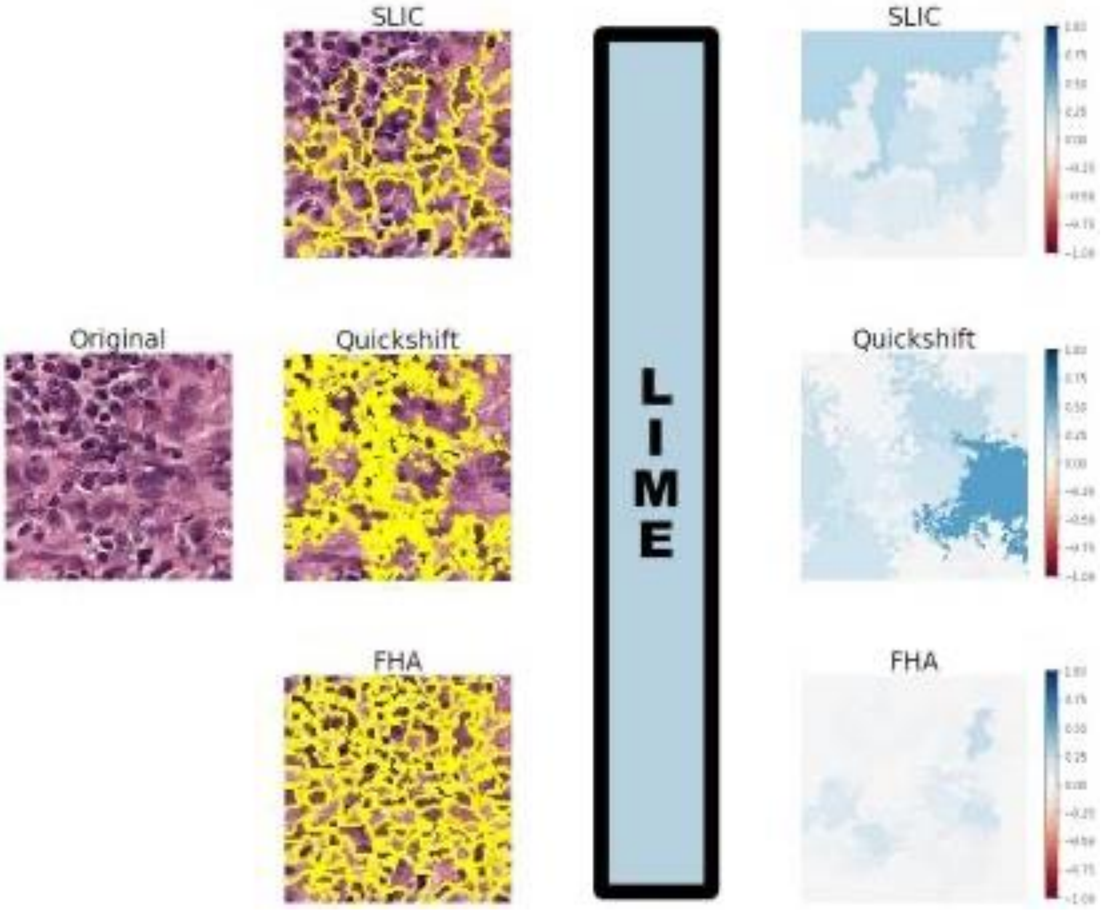


Figure 2.7: Explainable results produced by LIME when classifying Lymph Node Metastases. Here, different segmentation algorithms are compared (Palatnik de Sousa et al., 2019).

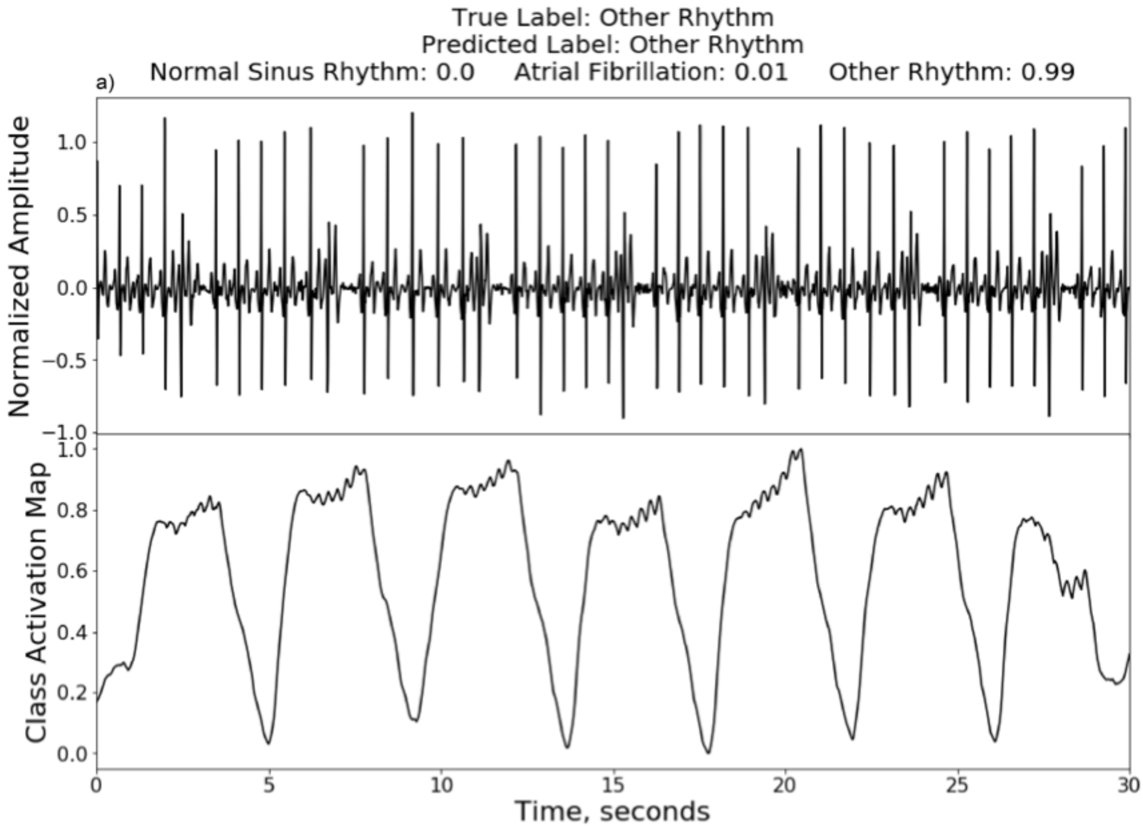


Figure 2.8: CAM results that reveal sections of an ECG signal relevant to predicting various rhythms (Goodfellow et al., 2018).

2.7.2 XAI for Emotion and Engagement Recognition

Research has looked at utilising XAI within affective computing. Affective computing is a field of computer science that focuses on systems and technologies that aim to recognise, interpret, and respond to human emotions. A recent review of XAI for affective computing looked into published research and techniques for global explanations, local explanations, pre-model, in-model and post-model XAI techniques (Cortiñas-Lorenzo & Lacey, 2023). Pre-model refers to techniques used to understand the data before the model is trained. In-model refers to XAI techniques that aim to understand how the data gets processed while the model is being trained. Post-model refers to techniques applied after the model has been trained, such as SHAP and LIME. A variety of techniques are used for post-hoc XAI, such as SHAP (Liew et al., 2021) and LIME (Chowdhury et al., 2021) for emotion recognition; however only one paper reviewed produced a multi-modal XAI system for emotion recognition (Pu et al., 2023) and none handled raw time series data. A multi-modal XAI method that can handle raw time series data would be important for affective computing as many of the well-established datasets contain multimodal time series data (Koelstra et al., 2012; Soleymani et al., 2012; Katsigiannis & Ramzan, 2018). The use of XAI for engagement detection/recognition is limited, and research seems to be in its preliminary stages, for example, one paper used an explainable decision tree model rather than a more advanced XAI technique to provide interpretability; however this came at the cost of model performance and accuracy (Rahman et al., 2022). As more advanced techniques, such as deep learning, are being developed and achieving state-of-the-art results, XAI methods need to be explored to understand these.

Time series data is commonly published in affective datasets and one of the most common and accessible types of data to research emotion recognition; therefore we will focus on some specific examples of XAI being used for models that handle time series data. Vielhaben et al., (2024) proposed a Layer-wise Relevance Propagation (LRP) method to provide visuals to explain univariate time series data. The method was demonstrated on ECG data, and sections of the signal the model looked at when making predictions were highlighted. They found that their model predominantly considered the mean of the signal but did also focus on the QRS complex (see 6.2.3.3.2). Ganeshkumar_et al., (2023) used grad-CAM to explain the outputs of a CNN that can classify multiple rhythm and morphological abnormalities. The activation maps in this study were used to ensure that the model was utilising known features that related to the ECG abnormalities. Neves_et al., (2021) aimed to make heartbeat classification more interpretable and then validated their method with ECG readers. LIME, SHAP and permutation sample importance methods were tested. LIME and permutation sample importance provided more relevant outputs than SHAP. It was found that visualisations of important segments of data would aid clinicians in their daily practice. There are very limited studies that specifically look at XAI for time series data for emotion recognition and of those that do, a large portion focus on video and EEG data. Torres et al., (2023) investigated the interpretability of deep learning algorithms that utilise EEG data to detect emotion. The features highlighted were consistent with known features in EEG literature.

2.8 Gaps in Literature

To this end, we have identified gaps in the literature that need to be explored in order to progress the understanding and development of at-home immersive virtual therapy. First of all, the ability to track patients' mental health outside of therapy sessions but throughout the course if therapy needs to be further explored (see 2.8.1). Whether VR is a good stimulus for emotion and engagement (both key components of therapy) and whether they can both be reliably detected within immersive virtual environments (see 2.8.2). In addition, explainable methods need to be researched to explain predictions that utilise multimodal time series data to clinicians (2.8.3).

2.8.1 Unobtrusive Mental Health Monitoring Throughout the Course of Psychotherapy of Patients Diagnosed with Depression.

A large body of research has delved into the field of mental health monitoring; however, the vast majority of these studies have focused on students or healthy adults. Limited studies have looked into utilising wearable sensors and machine learning to continuously track and predict the mental health of patients diagnosed with mental health diseases throughout the course of therapy. This needs to be understood as this population poses challenges that a healthy population does not, such as a skewed distribution of depression levels and higher levels of depression that are not present in a healthy population.

2.8.2 Are Immersive Experiences a Good Stimuli for Emotion and Engagement? How do they Compare to Non-immersive experiences? Can Emotion be Reliably Monitored Using ML?

The majority of well-established affective datasets were produced using non-immersive stimuli, the addition of an immersive VR-based affective dataset would open up more avenues of affective computing research. Due to the lack of immersive-based affective datasets, there is also a limited comparison of ML results directly comparing VR emotion recognition with the baseline results provided in established datasets. Another limitation of the current literature is the lack of analysis into the underlying importance of various physiological signal features in relation to emotion detection and how they correlate with what is currently understood in medical literature. This would contribute to more transparent ML systems that could be accepted by the medical community.

Engagement recognition is less researched than emotion recognition. To the best of our knowledge, whether various levels of engagement can be reliably detected within immersive virtual environments needs to be further researched. Similarly to emotion detection, there is also a lack of analysis in understanding the underlying feature importance of physiological data and its relation to engagement in immersive virtual environments.

2.8.3 Transparent and Interpretable Models for Emotion and Engagement Recognition, Specifically for Models Utilising Raw Time-series Signals.

Further developments need to be made in XAI for healthcare and affective computing. It has been stated that more research is specifically needed on XAI for 1d biosignals (Loh et al., 2022), such as

ECG and EEG, this research would improve the usefulness of wearable devices and patient monitoring. Along with this, the vast majority of post-model XAI techniques within affective computing are unimodal, leaving a gap for more multi-modal and modality-agnostic techniques to be researched (Cortiñas-Lorenzo & Lacey, 2023). This is especially important in affective computing, where the majority of emotional datasets consist of multi-modal time series. There is also limited research on the explain ability of eye-tracking time series data.

Chapter 3: Methodology for Algorithm Training, Validation, and Evaluation

3.1 Introduction

The purpose of this chapter is to outline the methodology used to train, validate, and evaluate the ML algorithms and techniques proposed in this thesis. There is no universally accepted validation method across the literature in emotion and engagement recognition (Ahmad & Khan, 2022). Inconsistencies such as using cross validation rather than subject wise cross validation can lead to variability in reported results, making direct comparisons between studies difficult (Bin Rafiq et al., 2020). By describing the methodology employed in this thesis, we aim to demonstrate the potential for variation in performance when using different validation techniques, and to emphasise the importance of employing robust evaluation frameworks that ensure generalisable, reliable results.

In addition to the methodological discussion, this chapter will also introduce the datasets used in this thesis, outlining their key characteristics, the modalities they encompass and the validation techniques used on them. Understanding the nature of the data is important, as it informs the selection of appropriate ML models and validation strategies.

The chapter is structured as follows: first, we provide an overview of training, validation and hyper-parameter tuning techniques commonly used in ML. The specific challenges that surround emotion and engagement Recognition are then discussed. Finally, we present the datasets and modalities used in this research.

3.2 Overview of training and validation techniques

3.2.1 Validation Techniques

In ML, proper validation techniques are crucial to accurately assess the performance and generalisability of models. This section outlines several widely used validation methods—K-fold cross-validation, random train-test splits, holdout set, and nested cross-validation—discussing their mechanisms, advantages, and limitations. Additionally, we will address the importance of stratified methods and the necessity of ensuring that data from the same participant is not included in both the training and validation sets, which is particularly important in emotion and engagement recognition tasks where individual variability can heavily influence model performance.

Random Train-Test Splits

The simplest method is the random train-test split. Here, the dataset is randomly divided into two subsets, typically with 70–80% of the data used for training and the remaining 20–30% for testing. The model is trained on the training set and evaluated on the test set. This method is popular for its simplicity and computational efficiency.

To address the challenges of imbalanced datasets, stratified train-test splits can be employed to ensure that both subsets maintain the same distribution of classes. This prevents the model from being biased towards the majority class and ensures a fair evaluation. However, in tasks such as emotion recognition, it is equally important to ensure that participant-specific data does not appear in both training and test sets. If the same individual's data is split across both sets, the model may learn participant-specific traits rather than generalisable patterns, leading to an overestimation of its generalisation capability (Bin Rafiq et al., 2020).

While random train-test splits are computationally efficient and straightforward, they can suffer from high variance in model performance depending on how the data is split, especially in small datasets. Without proper stratification or participant independence, this method may result in misleading performance estimates.

K-Fold Cross Validation

K-fold cross-validation is a resampling technique that partitions the data into K equally sized "folds." In this method, the model is trained on K-1 folds while the remaining fold is used for validation. This process is repeated K times, with a different fold serving as the validation set each time, and the final performance is averaged across all iterations. This approach ensures that every data point is used for both training and validation, which helps reduce bias and provides a more reliable estimate of model performance than a single train-test split. Typically, K is set to values such as 5 or 10, balancing computational efficiency and performance stability.

Leave-one-out cross-validation (LOO-CV) represents an extreme form of K-fold cross-validation where K equals the total number of data points. In each iteration, the model is trained on all data points except one, which is used for validation. This process is repeated N times, where N is the number of data points, ensuring that each data point serves as the validation set exactly once.

Care must be taken to ensure that data from a single participant does not appear in both the training and validation sets/folds. Failure to do so could result in models that overfit to individual-specific patterns rather than learning generalisable features. Therefore, subject-independent K-fold cross-validation, where each participant's data is entirely contained within either the training or validation set, is necessary to avoid inflated performance metrics that do not generalise well to new individuals.

This method has the advantage of using more data for training, reducing bias, and providing a robust evaluation of model performance. However, its computational demands, particularly on larger datasets, can be significant. Furthermore, if subject independence is not maintained, there is a risk of overestimating the model's ability to generalise.

Holdout Set

A holdout set is a commonly used technique in ML for evaluating model performance and generalisability. It involves dividing the available dataset into three distinct subsets: the training set, the validation set, and the holdout (or test) set. The model is trained on the training set, with hyperparameter tuning often performed using the validation set. The holdout set, however, remains completely unseen during both training and validation processes and is reserved exclusively for the final evaluation of the model.

The purpose of the holdout set is to provide an unbiased estimate of the model's performance on data it has never encountered before, simulating real-world scenarios where the model must make predictions on entirely new examples. Unlike cross-validation techniques, which rely on multiple iterations of training and testing on different data splits, the holdout method provides a single estimate of the model's generalisation ability. This approach is particularly useful when dealing with large datasets, where there is enough data to reserve a significant portion for the holdout set without compromising the quality of training.

Despite its simplicity, the holdout method has limitations. The model's performance estimate is subject to sampling variability, as the results depend on a single split of the data. If the holdout set is not representative of the overall data distribution, the evaluation may either overestimate or underestimate the model's performance. To mitigate this, larger datasets are often required for holdout validation to ensure that both the training and holdout sets capture sufficient variability in the data. In smaller datasets, other methods such as cross-validation are generally preferred due to their more reliable performance estimates.

In conclusion, while the holdout set is a straightforward and efficient method for evaluating model performance, especially in large datasets, care must be taken to ensure the split is representative and the dataset size is sufficient to avoid bias or misleading results.

Nested Cross Validation

Nested cross-validation is a more complex method used primarily when hyperparameter tuning is involved. It involves two levels of cross-validation: an outer loop that evaluates model performance and an inner loop for hyperparameter tuning. In this approach, the outer loop splits the data into training and testing sets, while the inner loop performs cross-validation within the training set to select optimal hyperparameters. Once the best configuration is identified, it is applied to the outer training set, and the model is evaluated on the outer test set. This process is repeated across all outer-loop folds, and the performance is averaged across iterations.

Nested cross-validation is particularly beneficial when comparing multiple models or algorithms, as it ensures that the hyperparameter tuning process does not lead to overfitting. By separating the hyperparameter tuning and final evaluation processes, this method prevents the model from implicitly learning information from the test set. However, as with other methods, it is essential to ensure stratification and that data from the same participant is not present in both training and test sets.

The primary advantage of nested cross-validation is its ability to provide an unbiased estimate of model generalisation, particularly when tuning hyperparameters. However, it is also one of the most computationally expensive validation techniques due to the two levels of cross-validation required.

Each of these validation techniques offers difffernt strengths and limitations, and the choice of method must be tailored to the nature of the dataset and the task at hand. Stratified methods are critical when dealing with imbalanced datasets, ensuring that all classes are represented appropriately in both the training and validation sets. Additionally, in the context of emotion and engagement recognition, it is essential to ensure that data from the same participant is not included in both the training and validation sets. Subject-independent validation is crucial to prevent overfitting to individual-specific traits and to provide a more accurate assessment of the model's ability to generalise to new participants. By considering these factors, we can ensure that the validation methods employed in this thesis produce robust, reliable, and generalisable results.

3.2.2 Normalising Data

Normalisation is an important step in ML preprocessing, as it aligns feature scales within a dataset, aiding model performance and stability. Algorithms that involve iterative learning, such as gradient-based methods, benefit from normalised data since features with a wide range of values can impede the model's convergence, potentially slowing down training or leading to instability. By standardising features to comparable ranges, normalisation also helps to reduce the risk of overfitting, as it prevents the model from weighting certain features disproportionately based on scale alone, which improves its generalisability to unseen data. Common methods, such as min-max scaling and z-score standardisation, are used to maintain this balance, allowing each feature to contribute appropriately to model learning. Throughout the analysis in this thesis, in order to normalise the data, the data was then standardised by subtracting the mean and scaling to unit variance.

$$\begin{gathered} z = (x - u)/s \\ x = the\ data\ needing\ to\ be\ normalised \\ u = the\ mean\ of\ all\ training\ samples \\ s = standard\ deviation\ of\ training\ samples \end{gathered} \tag{1}$$

3.2.3 Hyper-Parameter Tuning Techniques

Hyperparameter tuning is a critical aspect of optimising ML models, as it directly influences model performance. Hyperparameters, unlike model parameters, are not learned from the data but must be set before the training process. Finding the optimal combination of hyperparameters will improve a model's accuracy, generalisability, and robustness. However, improper tuning can lead to overfitting, where a model performs well on the training data but fails to generalise to unseen data. In this section, we discuss how to best tune hyperparameters while leveraging appropriate validation techniques to avoid overfitting.

Grid Search and Random Search

Two of the most common methods for hyperparameter tuning are grid search and random search. Grid search is an exhaustive method that explores every possible combination of hyperparameters within a predefined set of values. While this method can be thorough, it becomes computationally expensive, particularly when the number of hyperparameters is large. In contrast, random search selects random combinations of hyperparameters within the defined search space. While less exhaustive, random search is often more efficient, as it can explore a broader range of hyperparameter combinations without evaluating every single option (Yang & Shami, 2020).

Both grid search and random search should be combined with a robust validation technique, such as K-fold cross-validation, to avoid overfitting during hyperparameter tuning. By using K-fold cross-validation, we can ensure that the hyperparameters are evaluated on different subsets of the data, reducing the risk that the model overfits to any particular subset.

Nested Cross-Validation for Hyperparameter Tuning

While grid and random search can be useful for hyperparameter tuning, they can still result in overfitting if the same data is used for both tuning and final model evaluation. To address this issue, nested cross-validation is often recommended.

In nested cross-validation, the dataset is first split into training and test sets in the outer loop. The inner loop then performs K-fold cross-validation on the training set to tune the hyperparameters. Once the best hyperparameter combination is identified, the model is trained on the full outer training set with the chosen hyperparameters and evaluated on the outer test set. This process is repeated for each outer fold, ensuring a reliable performance estimate without overfitting.

Bayesian Optimisation and Automated Tuning

More advanced hyperparameter tuning techniques, such as Bayesian optimisation, offer a more efficient way to search the hyperparameter space. Bayesian optimization constructs a probabilistic model of the objective function and uses this model to select hyperparameters that are likely to improve model performance (Joy et al., 2016). This method balances exploration and exploitation by focusing on areas of the hyperparameter space that are more promising, rather than exhaustively searching or randomly sampling.

Automated hyperparameter tuning libraries, such as Hyperopt (Bergstra et al., 2015), often incorporate Bayesian optimisation and can be used in conjunction with validation techniques like K-fold cross-validation or nested cross-validation to minimise overfitting risks. These tools streamline the tuning process, reducing computational overhead while maintaining model robustness.

3.3 Challenges in Emotion and Engagement Recognition

Emotion and engagement recognition, particularly when using physiological signals, presents several unique challenges that are not always adequately addressed in traditional ML methodologies. A key difficulty lies in the high degree of individual variability inherent in physiological data. Factors such as age, health, baseline emotional states, and personal reactions to stimuli can cause significant variations between individuals (Ahmad & Khan, 2022), making it challenging for ML models to generalise well across diverse populations and possible to learn features related to a subject rather than the intended stimuli. Another significant challenge is the noise often present in physiological signals, due to sensor inaccuracies, environmental conditions, participant movement, or calibration which can obscure relevant patterns and lead to degraded model performance (Ahmad & Khan, 2022).

A risk in the emotion and engagement recognition literature is the overestimation of model performance due to inappropriate validation techniques. It is unclear in many studies that rely on training and validation procedures, such as K-fold cross-validation, whether data from the same participant is included in both the training and validation sets. This can lead to an inflated sense of a model's performance. For example, if data from a participant appears in both sets, the model may effectively "memorise" individual-specific characteristics, leading to artificially high accuracy, precision, and recall metrics. These inflated metrics do not reflect the model's ability to generalise to unseen participants, which is critical for real-world applications.

In this thesis we aim to demonstrate the impact of subject-dependent validation on performance overestimation. Studies that fail to incorporate subject-independent validation methods often report higher classification accuracies compared to those that implement leave-one-subject-out cross-validation or similar techniques.

Ensuring subject independence is essential not only for fair performance evaluation but also for developing models that are practical and reliable in real-world settings, such as clinical or therapeutic applications where new individuals are continuously introduced. By implementing these robust validation techniques, we can improve the reliability and generalisability of emotion and engagement recognition models, leading to more accurate and trustworthy systems.

3.4 Dataset, Modality and Validation Summaries

Here we present an overview of the datasets we use within this research, including validation techniques used when analysing the data. In some cases multiple different validation techniques are used for comparison. Randomised grid search will always be used for each fold when using nested cross validation and holdout sets.

Table 3.1: Overview of data and validation methodologies used throughout the thesis

Chapter	What is being predicted	Number of Participants	Number of data points	Modalities	Validation Methods
4	Levels of depression (PHQ-9)	17	160	Activity and sleep features recorded from a Fitbit (N features)	• Nested 10 fold cross validation • Holdout set • Nested leave one participant out cross validation
5	Emotion (4 quadrants of the CMA)	26	312	ECG, GSR, Eye Tracking	• Nested 10 fold cross validation • Holdout set • Nested leave one participant out cross validation
6	Emotion (4 quadrants of the CMA)	26	42,827	ECG, GSR, Eye Tracking	• Holdout set
7	Engagement	18	54 16,200*	EEG	• Nested leave one participant out cross validation

*54=complete samples, 16,200=5 second window samples

Chapter 4: Mental Health Tracking and Monitoring of Treatment-Resistant Depression (TRD) Patients Receiving Repetitive Transcranial Magnetic Stimulation (rTMS) Treatment

4.1 Introduction

In this chapter, we present a study investigating whether it is possible to use a commercially available wearable sensor, a Fitbit, to accurately track an individual's mental health status with a limited set of features. The focus is on using minimal features to track mental health in daily life throughout the course of therapy rather than using complex or specialised sensors. The goal is to determine if this approach is feasible, meaning that the results obtained from the machine learning (ML) models utilising the sensor data are consistent, accurate and clinically meaningful.

The data used in this study was collected from twenty-four patients diagnosed with clinical depression. These individuals underwent a specific treatment called repetitive transcranial magnetic stimulation (rTMS). rTMS is a non-invasive procedure that uses a magnetic field to stimulate nerve cells in the brain, and it is used to treat conditions such as depression, anxiety and chronic pain. Monitoring mental health states throughout and after therapy is essential as it can help evaluate the treatment's effectiveness and identify any potential relapses or setbacks. It has been found that when patients are monitored, treatment outcomes improve (Lambert et al., 2018).

Additionally, it can provide a sense of continuity and support for individuals at risk of mental health complications. One of the American Psychological Association guidelines on Evidence-Based Psychological Practice in Health Care is that "Psychologists aim to monitor the treatment process and clinical outcomes routinely" (Hamdi et al., 2021). It can also provide a way for individuals to stay engaged and motivated in their recovery.

ML models and sensors that can continuously track and analyse data could provide more support in monitoring mental health states. They could provide continuous monitoring even when a clinician isn't present, creating a complete picture of the mental health state through the course of therapy.

This study aims to explore ML analysis that could be used to track, monitor, and help clinicians understand depression severity over time of patients suffering from TRD, which can aid in diagnosing and treating the disorder. Our research objectives were as follows:

i) To analyse patient data and explore ML models and techniques to classify depression severity for a given day accurately.

ii) To analyse the features of the data, this analysis will highlight the most important aspects of the data for accurately predicting depression severity.

This project contributes to the overall thesis objective by exploring how an ML tool can be developed that can be used to track mental health states in daily life throughout the course of therapy, which can be used to monitor the progress of patients over time and to evaluate the effectiveness of the therapy. This relates to aim one of the complete thesis (see 1.2). Additionally, by analysing the features of the data, the project aims to understand the mechanisms that make this possible, which can inform future research and provide context for a future system that may need to be interpretable for clinicians using the data. This relates to aim five of the entire thesis (see 1.2). This chapter has been published in The Journal of Affective Disorders Reports (Griffiths et al., 2022).

4.2 Methodology

4.2.1 Data

This study recruited participants from an NHS trust's rTMS outpatient service who received rTMS treatment for depression between July 2019 and December 2020. Approval for the study was obtained from the Health Research Authority and Research Ethics Committee in the United Kingdom, and all participants provided informed consent. The inclusion criteria were: diagnosed with depression, aged 18 or over, and can read English. The exclusion criteria were: having a heart condition that a doctor determined negates participation; taking any photosensitive medicine; having epilepsy; bruise easily that prevents wearing of a wrist device; carpal tunnel syndrome that prevents wearing of a wrist device; and lack of capacity to consent.

Participants were given a Fitbit and received verbal and written instructions on how to use and access the Fitbit software application. Support was provided to let participants wear and use the Fitbit and its apps. Participants were required to download the Fitbit app to their smartphone, register with Fitbit, wear the device continually for the period of rTMS treatment (apart from when undertaking the rTMS) and charge it when required. At the outpatient visits, clinical staff reminded participants to wear the Fitbit and charge it. Activity and sleep data were collected using Fitbit.

A total of 24 participants who had TRD were included in this study; however, due to missing data (missing Fitbit sensor data and/or PHQ-9), only 17 participants have been included in the machine learning analysis where a total of 160 (M=9.41, SD= 8.43) days of data was used.

Activity Data

Physical activity can serve as an indicator of mental health. A lack of physical activity, either in the form of low levels of movement or prolonged sedentary behaviour, is often associated with increased severity of depression symptoms (Schuch et al., 2017). Engaging in physical activity is linked to reduced symptoms of depression in individuals diagnosed with depression (Teychenne et al., 2008). Individuals diagnosed with depression generally do less physical activity and more sedentary behaviour than the general population. In the general population, around 26% of adult activity levels can be labelled as sedentary (less than 5000 steps per day), 27% as low-level activity (5000 to 7500 steps per day), 17% as somewhat active (7500–9999 steps per day), 8% as active (10,000–12,499 steps per day), and 7% as highly physically active (12,500 steps and over per day) (TUDOR-LOCKE et al., 2009; Schuch et al., 2017). Activity data is captured by the Fitbit device. Using multiple measures of physical activity and heart rate, data is classified as sedentary, lightly active, moderately active, or very active using Fitbit's proprietary algorithms (Carpenter et al., 2021).

Sleep Data

Good sleep quality is associated with both mental and physical well-being. Sleeping for more than 9 hours or less than 7 hours per night can negatively affect an individual's health. (Watson et al., 2015). Around 35% of the general adult population gets less than 7 hours of sleep (CDC, 2022), which is associated with an increased risk of adult depression (Zhai et al., 2015). Conversely, having healthy levels of night-time sleep, around 7-9 hours, may decrease the risk of stress and mental illness by regulating the secretion of cortisol and other hormones. (Kumari et al., 2009, Meerlo et al., 2008). Sleep data is also captured by the Fitbit device. Using a combination of your movement and heart rate patterns, Fitbit estimates your sleep stages. The tracker or watch assumes the user is asleep when they haven't moved for about an hour (Fitbit help 2022).

PHQ-9

PHQ-9 was the measure used as our label for ML development. Measures of anxiety, depression and recovery were collected daily through PHQ-9 questionnaires. PHQ-9 is a self-report measure of depression; it has good sensitivity and specificity for major depression and good internal consistency (Kroenke et al., 2001). The PHQ-9 scores collected were transformed into a binary variable with labels of 'low to moderate' (PHQ-9 < 20) and 'high levels of depression' (PHQ-9 ≥ 20) (PHQ-9 Depression Test Questionnaire 2021).

Feature Extraction

ML analysis was carried out to explore the feasibility of classifying the level of depression severity using Fitbit data. Twelve features were collected from Fitbits, which include six physical activity and six sleep features as follows:

i) Activity features steps: minutes sedentary, minutes lightly active, minutes moderately active, minutes very active, and activity calories.

ii) Sleep features: minutes asleep, minutes awake, number of awakenings, and minutes of REM, light and deep sleep.

In addition, statistical features were extracted for our machine-learning analysis. These were the difference and the second order of difference for all features. The second order of difference is the difference between the first-order differences of a time series. It is used to detect changes in the rate of change of a time series, such as acceleration or deceleration. It is calculated by taking the difference between consecutive first-order differences. The reason for adding the difference and the second order of difference was to give contextual data outside the current day to the model. In total, 36 features were used and are displayed in Tables 4.1 and 4.2.

Table 4.1: List of Activity Features

Feature	Description
Steps	Number of steps taken that day
Minutes Sedentary	Data classified as sedentary, lightly active, moderately active, or very active by Fitbits proprietary algorithms (Carpenter et al., 2021).
Minutes Lightly Active	
Minutes Moderately Active	
Minutes Very Active	
Activity Calories	
Difference (for each feature above)	The difference in value between the current day and the previous day for a given feature
Second Order of Difference (for each feature above)	Difference between two consecutive first-order differences of the previous two days.

Table 4.2: List of Sleep Features

Feature	Description
Minutes Asleep	Total minutes of sleep
Minutes Awake	Total minutes awake during the night
Number of Awakenings	Number of times participants woke up during the night
Minutes of REM Sleep	Minutes spent in the REM stage of sleep
Minutes of Light Sleep	Minutes spent in the light stage of sleep
Minutes of Deep Sleep	Minutes spent in the deep stage of sleep
Difference (for each feature above)	
Second Order of Difference (for each feature above)	

Data Exploration

An overview of the data and its distribution can be seen in Figures 4.1, 4.2, and 4.3. The distribution of PHQ-9 scores clearly highlights that our dataset is skewed towards severe levels of depression. For the feature data, there are clear bell curves around zero for the change in feature and second order of difference in features. This suggests there would be low variation within individual participant's non-previous features. This would mean that the reason for the less uniform distributions seen in the non-previous features could be due to differences from participant to participant.

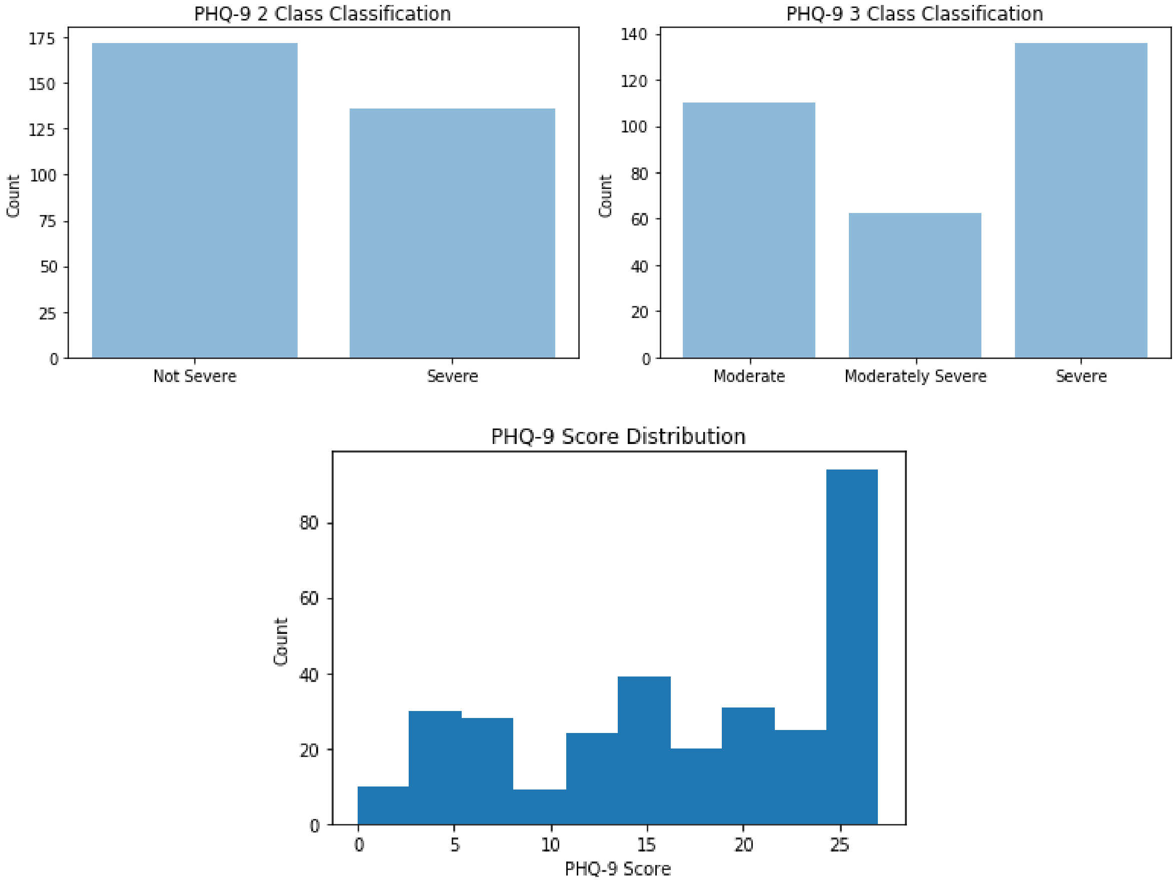


Figure 4.1: PHQ-9 Distributions

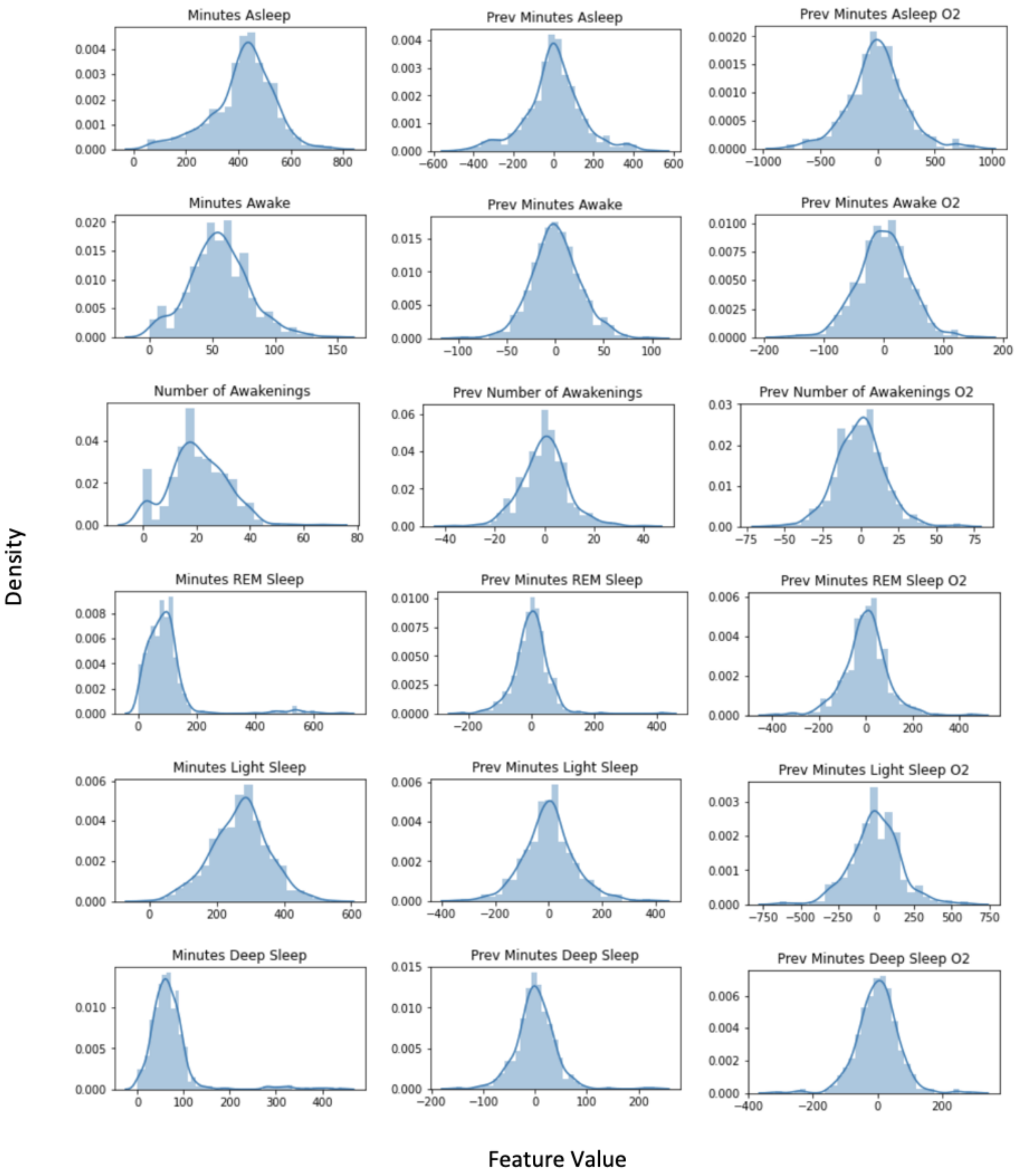


Figure 4.2: Sleep Feature Distributions

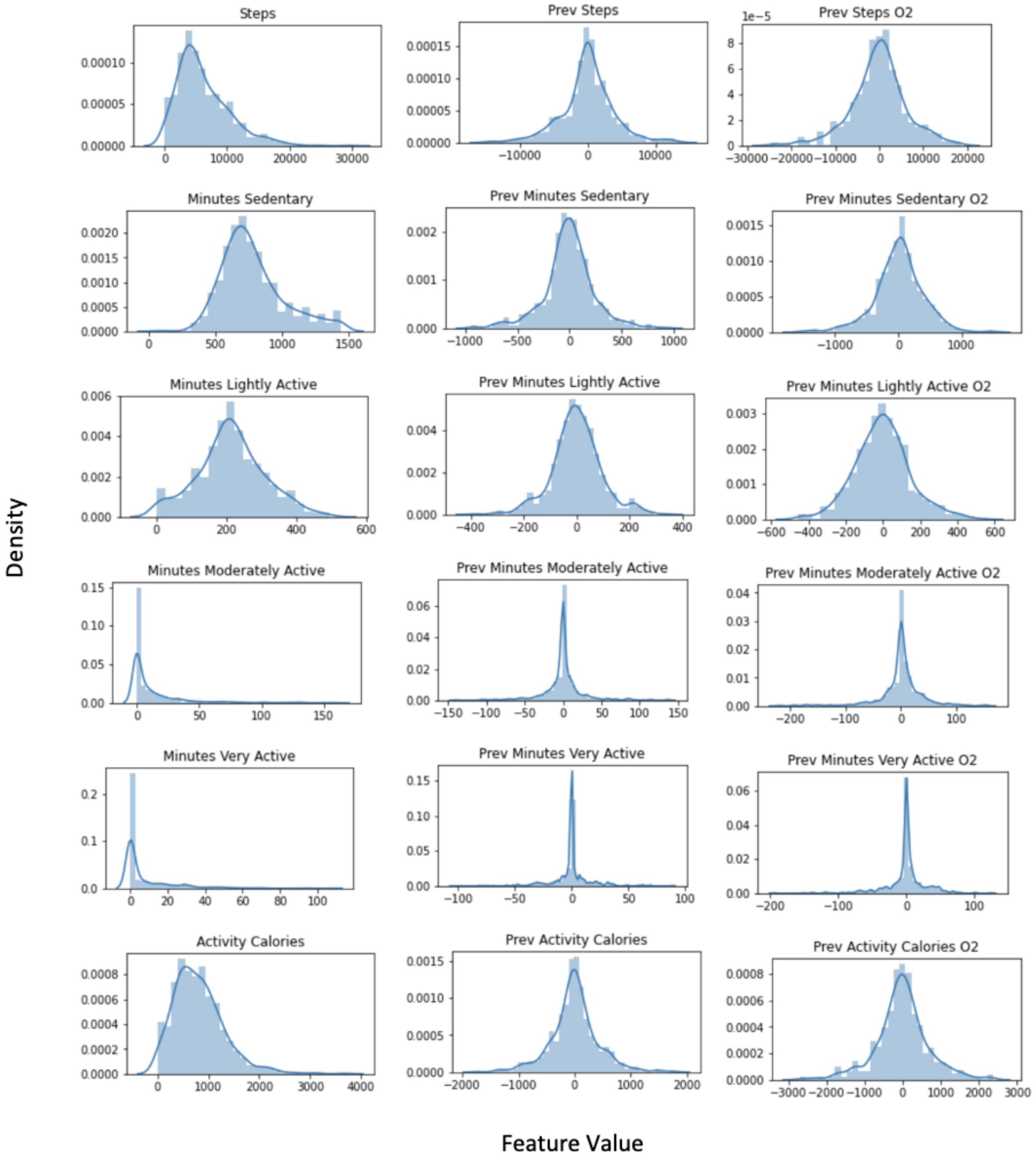


Figure 4.3: Activity Feature Distributions

4.2.2 Model Development

4.2.2.1 Pilot

Five models were selected to be tested in the pilot trial; these were Support Vector Machines (SVM), K-nearest Neighbour (KNN), Logistic Regression (LR), Decision Trees (DT) and Random Forest (RF)—the pilot trial aimed to evaluate the best algorithm to evaluate further with different validation methods. 10-fold nested cross validation was performed on this data with each algorithm. In each nested fold, hyper parameters were tuned on the inner loop before validating on the outer loop. Therefore all results presented in the pilot trial have been obtained on tuned models. The hyper parameters were tuned using randomised grid search (see section 3.2.2).

SVMs are popular for many classification and regression tasks because they are effective and have relatively few hyperparameters to tune. However, they can be computationally expensive to train, especially for large datasets, and there may be better choices for problems with very large numbers of features or instances. Since this data set does not have many features or instances, SVMs were used in the initial model selection phase. KNN is a strong candidate for this problem as it can be used with a small amount of data and can handle missing data. LR is a simple and effective tool for classification tasks with relatively few hyperparameters to tune, and it can work well with a wide range of datasets. However, it may not be as powerful as more complex models, such as SVM or RF, for tasks with highly non-linear patterns or many features. Due to their interpretability, DT are commonly used in medical diagnosis applications (Azar & El-Metwally, 2012). However, our small dataset may not be suitable because DT can be prone to overfitting; a RF may be more robust and generalise better when working with small datasets. RF are widespread for many machine-learning tasks because they are relatively easy to implement and often perform well without requiring extensive hyperparameter tuning. They are also resistant to overfitting, which makes them a good choice for datasets with a small number of samples or a large number of features. Please see section 2.5 for more information on each algorithm.

Pilot Results

The results of our pilot trial, including the mean accuracy and standard deviation, are shown in table 4.4 and figure 4.4. When selecting the algorithm to proceed with, performance metrics, interpretability, computational requirements, model complexity, feature importance methods and support from the literature were considered. First, DT and SVM were dismissed because their performance was considerably lower than that of LR, KNN and RF algorithms. LR, KNN and RF all have similar accuracy.

Interpretability was a significant factor in algorithm selection. Not only would an interpretable model be important from a usefulness standpoint, but it is critical from an ethical standpoint, considering treatment decisions could be made taking into account the models' results. Feature importance methods that can be used on these algorithms were also taken into account in the decision-making process; it is sensible to hypothesise that there is a multicollinearity of features, so the feature importance method is needed to handle that. Computational complexity was also taken into account. RF tends to have slightly higher computational requirements and model complexity; however, with the dataset being small and having a relatively small amount of features, the difference between the two will be minimal for this task and will not be discussed further. The final consideration in our algorithm selection was the support throughout the literature when used in similar tasks.

Considering the discussion above, RF was selected to proceed with mainly due to the performance in the pilot trial and the feature importance methods that can be used in conjunction with it. LR may have had a slight advantage in interpretability, but not enough to outweigh the benefits RF provides in feature importance and pilot classification results. In Table 4.3, we present the discussions for each of the characteristics we considered in the decision-making process (feature importance methods, interpretability and support in literature).

Table 4.3: Model Comparison for Model Selection

Characteristic	**Logistic Regression**	**Random Forests**	**KNN**
Feature Importance	Coefficients can be used to highlight feature importance; however, can be affected by multicollinearity.	This can be achieved using a tree-based method; multicollinearity is generally not a significant issue when assessing the importance of features in tree-based models because these models are based on recursive partitioning and can handle correlated predictors.	KNN does not provide inherent feature importance measures because it is a non-parametric, instance-based learning method. Feature scaling, however, can affect distances and influence prediction.
Interpretability	Generally considered a very interpretable model. This is because the model produces coefficients for each input feature, which can be used to understand how each feature contributes to the final predicted probability.	Generally considered less interpretable than LR models, when many DTs are combined into an RF, the resulting model becomes more complex and challenging to interpret. Despite this complexity, some techniques can still be used to understand an RF model's decision-making process	Considered less interpretable; KNN works by comparing the distance between data points, which does not provide coefficients or intuitive weights. Only the chosen distance metric is visible to users.
Support in literature	Chikersal et al., 2021; classified depression using smartphone usage features Saeb et al., 2015; used mobile phone	Wahle et al., 2016; used GPS and smartphone usage to detect depression in the general population (Pedrelli et al., 2020): wrist-based wearable sensors that measured electrodermal	Kılıç et al., 2022; used sleep and activity data to predict ‘subjective well being'. Multiple algorithms were tested and KNN performed comparatively well in all scoring metrics and was

sensors to predict depression severity Wang et al., 2018; used mobile and wearable sensors to predict whether a student is depressed or not week to week.	activity, skin temperature, and heart rate, along with mobile usage data, were used to predict depression biweekly.	the fastest algorithm to train.

Table 4.4: Pilot Results

Model	Mean Accuracy	Standard Deviation
Logistic Regression (LR)	0.71	0.16
Random Forest (RF)	0.73	0.15
Decision Tree (DT)	0.63	0.12
Support Vector Machine (SVM)	0.60	0.08
K Nearest Neighbour (KNN)	0.72	0.14

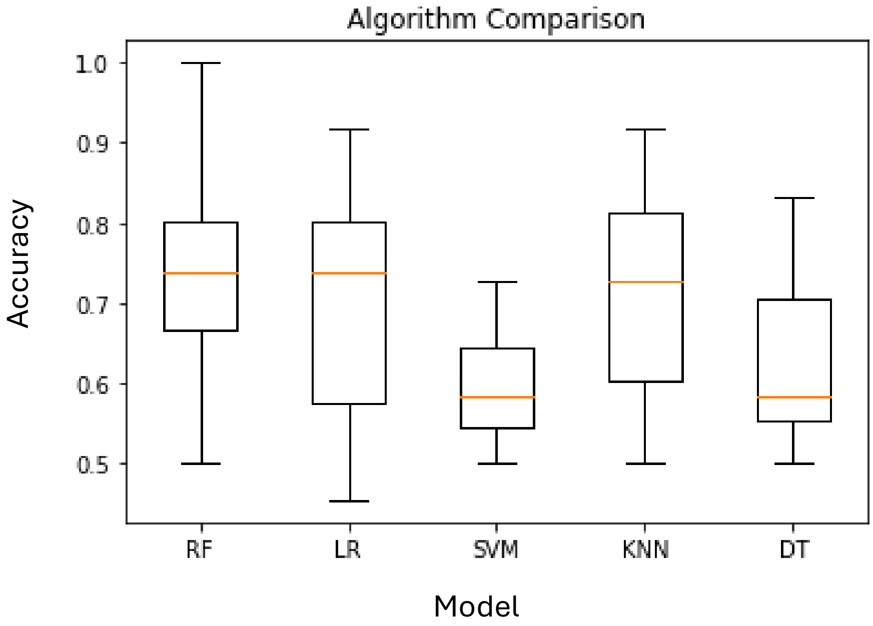


Figure 4.4: Pilot Results

4.2.2.2 Model Evaluation

Three different validation methods were used on this data, a holdout set, ten-fold nested cross validation and leave one participant out nested cross validation. For the holdout set, ten-fold cross-validation was performed on 90% of the data to train and validate our selected model, leaving 10% for unseen testing. For ten-fold cross nested cross validation, the complete dataset was randomly split, meaning that data from the same participant may appear in both the training and validation sets. The final validation method, leave on participant out nested cross validation ensured data from a single participant was not in both the training and validation data set.

Hyper parameters were tuned in the nested cross validation using randomised grid search. the list of hyperparameters and values considered can be found in table 4.5.

Table 4.5: List Hyperparameters and Values Considered

Hyper Parameter	Values	Description
minimum sample leaf	1, 2, 4	The minimum number of samples required to be at a leaf node.
minimum sample split	2, 5, 10	The minimum number of samples required to split an internal node
n estimators	200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000	The number of trees in the forest.

Three different classification tasks were tested, severity classification, PHQ-9 change from previous day and PHQ-9 change from baseline assessment. One of the essential components of evidence-based practice in mental health is regularly monitoring client progress during therapy. (APA Presidential Task Force on Evidence-Based Practice 2006; Dozois et al. 2014). Consistently, research has found that collecting progress data using standardised rating scales on a session-by-session basis and utilising the feedback to make clinical decisions can lead to improved outcomes and reduced risk of treatment failure, especially for clients at a higher risk. (Bickman et al. 2011; Lambert et al. 2003). Despite the evidence that monitoring and feedback can lead to better client outcomes, available data suggest that this practice is not commonly implemented in clinical settings. (Ionita and Fitzpatrick 2014). One reason for the lack of implementation of monitoring and feedback in clinical settings is the issue of clients' willingness to complete the necessary measures. (Kotte et al. 2016)

Resource constraints and the additional time and paperwork required for monitoring and feedback implementation are other reasons that can contribute to the limited adoption of this practice in clinical settings. (Gleacher et al. 2016). Therefore, we looked at three different classification tasks that could aid in solving this problem;

i) Severity classification: A measure of depression severity relating to PHQ-9 ratings (see 3.2.1.3 for a breakdown of the low, medium and high bands). Having a patient's daily severity without input from the patient or clinician would mitigate these reasons. Severity classification would allow constant understanding of the patient's current level of depression.

ii) Change from the previous day: Change from the previous day classifies if, on the current day, the patient's PHQ-9 scores are lower, equal, or higher than the previous day's results. The change from the previous day provides further context where, for example, a patient's severity level may not have changed, but there is still a change from the previous day. This would be useful in viewing overall trends and making monitoring how patients react to treatment easier.

iii) Change from baseline: This works and provides the same benefits as the change from the previous day; however, we are looking at the change from the baseline tests from the beginning of treatment instead.

In addition to the classification tasks, we investigated feature importance. To understand the importance of the different feature sets in the classification task, we individually analysed the activity and sleep feature sets before analysing the combined feature sets (which include activity, sleep and statistical features).

As an RF is being used to analyse this data, a tree-based feature importance method would provide a comprehensive understanding of the most important features of our model. The feature importance we provided was based on the mean decrease in impurity within each tree. Our random forest criterion function is Gini (a measure of the probability of a random sample being classified incorrectly), and the feature importance measures the importance of each feature by the decrease in the Gini impurity caused by splits on that feature. Features that result in larger decreases in Gini impurity are considered more important and are given a higher importance score. The scikit learn implementation of this was used (Pedregosa et al., 2011).

Based on the feature importance, we then trained new models with various numbers of features (5, 10, 15, 25 and 30 features). Ten-fold nested cross-validation was used random forest model (tuned on each fold) to find the optimal number of features. We assessed model performance with ten selected features using the three validation methods used in the previous tasks. For this analysis, only 9 participants had enough data points (minimum of 4 and maximum of 32) to be included in this model. All results were then compared.

4.3 Results

4.3.1 Classification Results

Depression Severity Classification

The results of our machine learning analysis are summarised in Table 4.6. Results on activity, sleep, and combined feature sets provided insight into each feature set's importance and the differences in performance. Specifically, our model can classify two levels of depression severity using activity, sleep, and combined feature sets with an accuracy of 65%, 69, and 70%, respectively when leave one participant out (LOPO) nested cross validation was used. The holdout set produced the highest overall accuracy when using combined features but the worst for activity and sleep features. 10-fold cross validation performed better that LOPO cross validation which is to be expected, however the differences in performance were not major.

Table 4.6: Severity Classification Results

Feature Set	Precision	Recall	F1	Accuracy
Activity features holdout set	0.55	0.55	0.55	0.55
Activity features nested 10 fold CV	0.70 (SD=0.12)	0.69 (SD=0.01)	0.70 (SD=0.07)	0.66 (SD=0.11)
Activity features nested LOPO CV	0.70 (SD=0.27)	0.65 (SD=0.20)	0.66 (SD=0.22)	0.65 (SD=0.20)
Sleep features holdout set	0.69	0.69	0.69	0.69
Sleep features nested 10 fold CV	0.74 (SD=0.09)	0.81 (SD=0.12)	0.75 (SD=0.08)	0.68 (SD=0.05)
Sleep features nested LOPO CV	0.88 (SD=0.22)	0.69 (SD=0.22)	0.69 (SD=0.22)	0.69 (SD=0.21)
Combined features (activity + sleep + statistical features) holdout set	0.75	0.75	0.75	0.75
Combined features (activity + sleep + statistical features) nested 10 fold CV	0.78 (SD=0.13)	0.84 (SD=0.21)	0.78 (SD=0.11)	0.73 (SD=0.15)
Combined features (activity + sleep + statistical features) nested LOPO CV	0.77 (SD=0.12)	0.66 (SD=0.12)	0.68 (SD=0.17)	0.70 (SD=0.12)

Change from the Previous Day and Baseline Classifications

In this analysis PHQ-9 change from previous day and baseline were being classified. All features were included in this analysis. Both tasks were 3 class classification tasks consisting of lower, equal, or higher than the previous day/baseline. Our model can classify change from baseline at 62% accuracy and the change from the previous day at 65% accuracy when LOPO nested cross validation was used. Results from the other validation methods outperform LOPO, with the best performing validation method being the holdout set achieving 74% and 84% accuracy when classifying change from baseline and previous day respectively.

Table 4.7: Change from the Previous Day and Baseline Results

Classification Task	Precision	Recall	F1	Accuracy
Change from baseline holdout	0.71	0.74	0.72	0.74
Change from baseline 10 fold CV	0.68 (SD=0.14)	0.72 (SD=0.18)	0.70 (SD=0.14)	0.69 (SD=0.15)
Change from baseline nested LOPO	0.59 (SD=0.25)	0.65 (SD=0.32)	0.64 (SD=0.20)	0.62 (SD=0.21)
Change from the previous day holdout	0.83	0.83	0.83	0.83
Change from the previous day 10 fold CV	1.00 (SD=0.30)	0.07 (SD=0.13)	0.05 (SD=0.15)	0.75 (SD=0.04)
Change from previous day nested LOPO	0.64 (SD=0.32)	0.65 (SD=0.25)	0.64 (SD=0.25)	0.65 (SD=0.25)

4.3.2 Feature Importance

Understanding the feature's importance is valuable as it can provide useful insights into future work to improve classification performance, reduce model complexity, and improve training and running speed. In addition, feature importance could provide real-world insights into what seems to have the most impact on depression severity.

We ranked the importance of our ten selected features based on the 'importance score' (see Table 4.8). Most features selected were sleep-related, with the most important being the number of awakenings during the night.

Table 4.8: Feature Importance Results

Feature	Importance
Number of Awakenings	0.217
Minutes Lightly Active	0.191
Minutes Light Sleep 2nd Order of Difference	0.12
Minutes Light Sleep	0.11
Minutes Very Active	0.103
Minutes Deep Sleep	0.086
Minutes Awake	0.061
Minutes REM Sleep	0.059
Minutes Sedentary	0.031
Minutes Moderately Active	0.021

4.3.3 Classification with Selected Features

In addition to the original classification and feature importance and selection, we also investigated the number of features to use. This analysis is useful for reducing model complexity but could also possibly highlight features that may have a negative impact on the model. To find the optimal number of features, we used the feature importance scores to select features and examine if the model performance could be improved. Ten-fold cross-validation was used on a tuned random forest model to find the optimal number of features. The top 5, 10, 15, 20, 25 and 30 features were tested; these results can be found in Figure 4.5, showing that the model with ten selected features achieved a high performance. All three validation methods used throughout this chapter were used again for this task. LOPO nested cross validation achieved 72% accuracy. The best performing validation method was 10-fold nested cross validation achieving 77% accuracy.

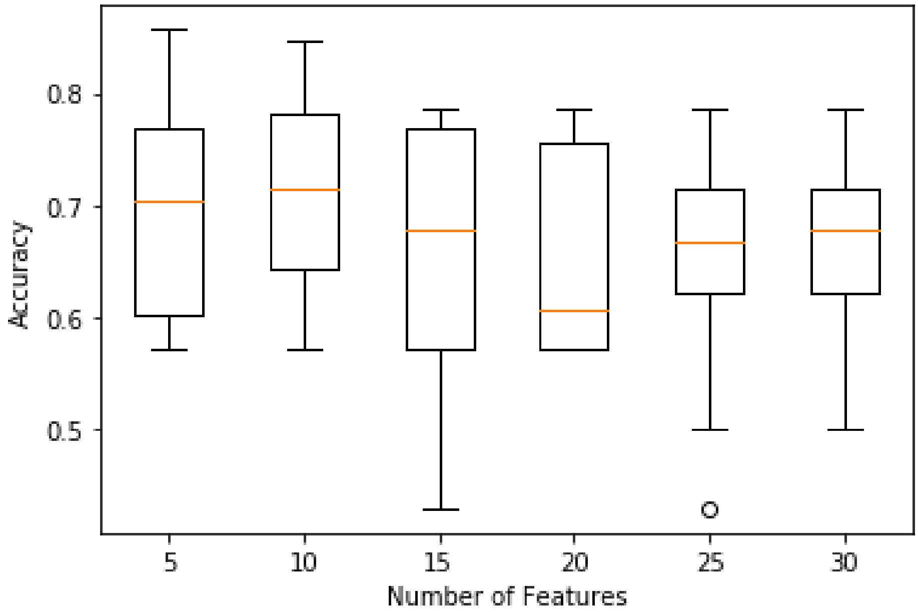


Figure 4.5: Number of Feature Results Comparison

Table 4.9: Selected Features Results

Feature Set	Precision	Recall	F1	Accuracy
All features Holdout	0.75	0.75	0.75	0.75
All features nested 10 fold cv	0.78 (SD=0.13)	0.84 (SD=0.21)	0.78 (SD=0.11)	0.73 (SD=0.15)
All features nested LOPO	0.77 (SD=0.12)	0.66 (SD=0.12)	0.68 (SD=0.17)	0.66 (SD=0.12)
Ten selected features Holdout	0.82	0.81	0.81	0.81
Ten Selected Features nested 10 fold CV	0.78 (SD=0.18)	0.72 (SD=0.21)	0.75 (SD=0.16)	0.77 (SD=0.13)
Ten selected features nested LOPO	0.71 (SD=0.22)	0.76 (SD=0.56)	0.74 (SD=0.20)	0.72 (SD=0.19)

4.4 Discussion

4.4.1 Research Objective I: Exploration of ML Models and Techniques that can Accurately Classify Depression Severity

Our machine learning model can classify two levels of depression severity using the activity, sleep and combined features; utilising the ten selected features, it achieved 72% accuracy (using LOPO nested cross validation). There are no direct comparisons to our machine learning results, as our dataset is unique because it consists of TRD participants. Prior to this research, no other study has investigated whether machine learning can be used to track the depression levels of TRD patients throughout the course of therapy. Classifying TRD patients' depression levels is unique because we are classifying slight differences in depression at the extremely high end of the scale. Successfully analysing this suggests that models can be developed to track populations of people that would be considered outliers among the healthy population. This research also shows the viability of ML methods to track and monitor depression in a real-world clinical setting. Other research primarily focuses on participants from the general population, using physiological, social media and smartphone data (Acharya et al., 2015; Intarasirisawat et al., 2020; Wang et al., 2013). A study classified depression using smartphone usage features; 138 university students were recruited to develop a model to classify if they had depressive symptoms post-semester; an accuracy of 83.3% was achieved (Chikersal et al., 2021). Another study recruited adults from the general population and used GPS and smartphone usage to detect depression; using data from 28 participants, the model achieved 86.6% accuracy in discriminating between participants below or equal to 5 and above 5 PHQ-9 (Wahle et al., 2016). A summary of the comparisons to other similar studies is presented in Table 4.10. In our analysis, the majority of features selected were sleep-related, the most important being the number of awakenings during the night. This shows that a high depression level correlates with disturbances in sleep, which is consistent with the literature (Zhai et al., 2015).

Among the three validation methods we used, ten-fold cross-validation demonstrated the highest performance. While this result was anticipated, it underscores how this method can overestimate a model's generalisability. In contrast, the holdout set yielded the most variable results across tests, reflecting its sensitivity to the randomness of data sampling. This inconsistency suggests that, with a dataset of this size, the holdout method may not provide an accurate estimation of model performance.

It should be noted that our study is limited by the size and specificity of our sample population, which may impact the generalisability of the model. Smaller datasets like ours can lead to potential overfitting, particularly when using certain validation methods, as highlighted by the variability in results from the holdout method. Future studies should validate these findings on larger, more diverse TRD populations and explore additional feature sets to refine model accuracy and robustness further.

Table 4.10: Comparison of Results to Literature

Reference	Participants	Data	Classification	Accuracy
This paper	17 patients diagnosed with TRD	Fitness tracker	Detection of depression severity daily	Accuracy=81%
(Chikersal et al., 2021)	138 university students	Smartphone sensors, usage statistics and fitness tracker	detection of depressive symptoms after a semester	Accuracy=83.3%
(Saeb et al., 2015)	28 adults	Location	Depression at the end of 2 weeks	Accuracy=86.5%
(Wahle et al., 2016)	36 adults	Smartphone sensors	Depression biweekly	Accuracy=61.5%
(Farhan et al., 2016)	79 university-age people	Location	Clinical depression biweekly	F1=0.82

(Canzian and Musolesi, 2015)	28 adults	Location	Detecting depression over different periods of time and in advance	Sensitivity=0.71
				Specificity=0.87

(Wang et al., 2018)	68 university students	Smartphone sensors	Depression weekly	F1=0.75
(Mehrotra et al., 2018)	28 adults	location	Detecting depression over different periods of time	Sensitivity=0.77

(Pedrelli et al., 2020)	31 Adults with Major Depressive Disorder (MDD)	Mobile data and wearable sensors	Depression biweekly	MAE=3.88

4.4.2 Research Objective II: Feature Importance Analysis

Six of the selected features are sleep-related. This fact and the sleep features-only model's results outperform the activity features-only model throughout all validation methods, indicating that sleep is a primary factor when building depression monitoring ML models. Specifically, the Periods of light sleep and awakenings significantly affect the model. In literature, irregular sleep has also been linked to poor mental health (Ghandeharioun et al., 2017). All other studies in our results comparison table use either smartphone sensors, usage, location, or all three and achieve good results. It would be reasonable to expect a better model to be built if the phone and location features could be used alongside activity and sleep.

Feature selection suggests that our model performs better with relatively low (ten) features; this could be because our model was trained on a small dataset and could be prone to overfitting, which feature reduction would have helped with. With this point in mind, future research that may include more data could benefit from the complete feature set. The increased simplicity is a benefit of using fewer features, particularly for this use case. Models and their predictions would be easier to interpret and explain. For example, feature correlations could be further investigated; for example, there is a clear negative correlation between the more sedentary minutes and the minutes spent asleep (Figure 4.6); this information interpreted by a doctor could be used to guide patients. We can also compare our models' predictions to information found throughout the literature; for example, engaging in physical activity is linked to reduced symptoms of depression in individuals diagnosed with depression (Teychenne et al., 2008), and a lack of physical activity, either in the form of low levels of movement or prolonged sedentary behaviour, is often associated with increased severity of depression symptoms (Schuch et al., 2017). These statements can be clearly seen if we plot participants' very active and sedentary minutes (both of which were in the ten selected features) with their depression level highlighted (Figure 4.7). The chart clearly shows that less sedentary minutes tend to mean more very active minutes and lower PHQ-9 levels.

COVID-19 restrictions on social activity and movement applied during the study may have affected the activity levels. Therefore, this may be a reason activity features seem to be less critical. Further study outside of COVID-19 regulations will be needed to verify this. Furthermore, this data may differ from the data collected now; therefore, additional model tuning and development may be needed to make a reliable model for daily life without COVID-19 restrictions.

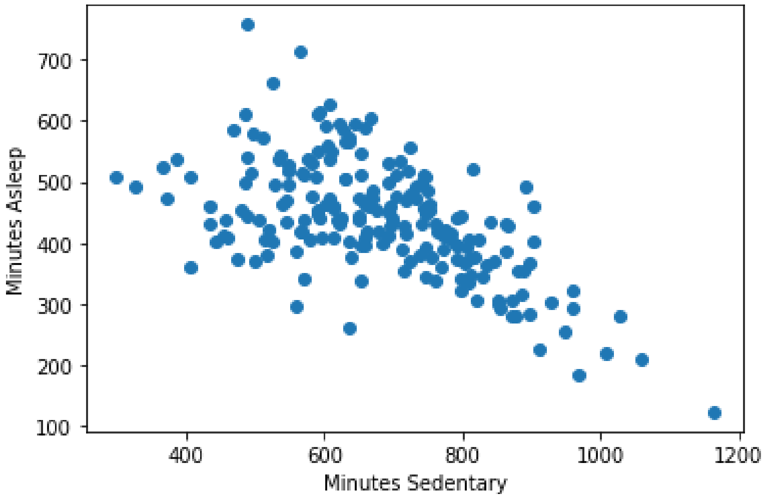


Figure 4.6: Minutes Asleep and Minutes Sedentary Scatter Plot

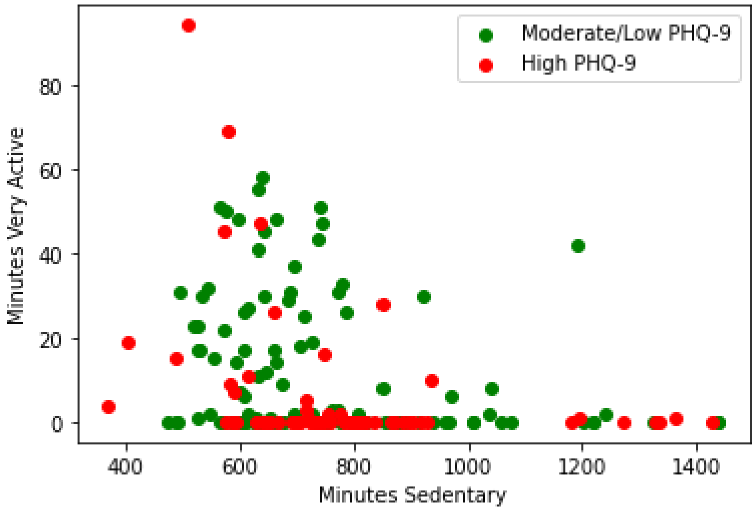


Figure 4.7: Minutes Very Active and Minutes Sedentary with Corresponding PHQ-9 Levels Scatter Plot

Chapter 5: Machine Learning based Emotion Recognition During Virtual Reality Experiences

5.1 Introduction

Emotions can be described as subjective experiences that involve psychological and physiological reactions and responses (Hockenbury, 2010). They are a crucial aspect in our lives, constituting an essential part of decision-making, social interaction, perception, memory, learning and creativity (Tripathi et al., 2017, Zhang et al., 2015). Emotions can be expressed through verbal or non-verbal cues (i.e. facial expressions, gestures, voice, etc.) (Argyle et al., 1970). A plethora of literature has explored the use of technology-based emotion assessment and recognition modalities to further understand human experiences in many domains, such as psychology (Iffland et al., 2010; Kosunen et al., 2016; Nuske et al., 2014), healthcare (Kumar et al., 2019), education (Lane & D'Mello, 2018), advertisement (Grigorovici & Constantin, 2004) and tourism (Mohd et al., 2019).

Assessing and evaluating emotional states are of specific importance in the human-computer interaction (HCI) community (Zucco et al., 2017); understanding emotional responses can help researchers design better user experiences, hence improving the usability, acceptability, and accessibility of technologies. For instance, HCI researchers have used physiological responses to enhance game design and experience (Lin & Imamiya, 2006), evaluate website design and usability (Qu et al., 2017), and examine the acceptability of assisted interactions using smart conversational agents (such as Apple Siri and Amazon Alexa) (Lee et al., 2020). Furthermore, emotion recognition is paramount in affective computing research; enabling intelligent systems to recognise, infer and interpret human reactions to improve user experiences as well as enhance their outcomes (Kosunen et al., 2016).

Researchers have utilised an array of sensors to evaluate and monitor human behavioural and physiological biomarkers such as electrocardiogram (ECG) (Nardelli et al., 2015; Zhang et al., 2015), galvanic skin response (GSR) (Macedonio et al., 2007; Valenza et al., 2012), electroencephalogram (EEG) (Jalilifard et al., 2016; Kosunen et al., 2016; Tripathi et al., 2017) and eye tracking (Salvucci & Goldberg, 2000). In such studies, emotional elicitation is normally induced by exposing subjects to emotional experiences, situations, or stimuli. To boost the advances in HCI in general and affective computing in particular, many researchers have generated and shared annotated multimodal affective datasets in which other researchers can easily access data that are time-consuming to generate otherwise and directly develop their methodologies and test their hypotheses or algorithms. There are currently several popular affective datasets that combine psychological, behavioural, and physiological data, such as MAHNOB-HCI (Soleymani et al., 2012), DECAF (Abadi et al., 2015), DEAP (Koelstra et al., 2012), DREAMER (Katsigiannis & Ramzan, 2018) and WESAD (Schmidt et al., 2018). In all these datasets, emotional responses were triggered using non-immersive modalities such as video clips (Abadi et al., 2015; Katsigiannis & Ramzan, 2018; Koelstra et al., 2012; Soleymani et al., 2012) or audio (Panda et al., 2020). There is however, limited research into utilising physiological signals for emotion recognition in immersive environments.

Understanding emotional responses within virtual environments could benefit healthcare, particularly immersive virtual therapy. Virtual reality (VR) exposure therapy is a form of psychotherapy that uses VR to treat mental health conditions such as anxiety disorders, phobias, and post-traumatic stress disorder (PTSD) (Parsons & Rizzo, 2008; Eshuis et al., 2021; Difede & Hoffman, 2002; Albakri et al., 2022). In VR exposure therapy, patients are exposed to simulations of feared environments or situations in a controlled and safe setting, allowing them to practice coping strategies and improve their emotional regulation (see more in 2.3 and 2.4). Clinicians need to be able to understand a patient's emotional state during therapy for several reasons, including tailoring treatment and assessment (see 2.4 for further discussion). Overall, the ability to understand and respond to a patient's emotional state is a crucial component of effective therapy, as it enables clinicians to provide personalised, evidence-based care.

This study aimed to create and validate a dataset where users' emotions were elicited using virtual environments delivered via a VR headset. Machine learning (ML) analysis was utilised to demonstrate the application of the dataset in emotion recognition research. The results presented in this chapter are compared with other well-known and established emotion recognition datasets and the difference between immersive and non-immersive stimuli was investigated. Feature importance was also performed on the feature set. Our research objectives were as follows:

I) Understand whether VR is a reliable stimulus for emotion recognition. This relates to the overall thesis objective 2 (see 1.2)

II) Investigate whether robust ML models can be developed to recognise emotion within virtual environments and how it compares to ML models that are built using non-immersive stimuli. This relates to the overall thesis objective 3 (see 1.2)

III) Investigate the specific features of data that indicate emotion and compare these to affective computing and medical literature to make the ML models more transparent. This relates to overall thesis objective 5 (see 1.2)

5.2 Methodology

5.2.1 Affective Responses

The circumplex model of affects (CMA) (Russell, 1980) is a widely used model that interprets emotional effects as a continuum of highly interrelated states. The CMA is a bi-dimensional model, where the valence dimension ranges from positive (i.e. happy, relaxed) to negative (i.e. nervous, sad), and the arousal dimension ranges from low arousal (i.e. calm, depressed) to high arousal (i.e. tense, excited) (see Figure 2.3). For more information, see section 2.4. In this research, we aimed to evoke and detect the four quadrants of the CMA.

5.2.2 Stimuli Selection

The selection of effective stimuli is essential for eliciting appropriate affective responses; therefore, the selection process underwent several stages. The study aimed to select three virtual environments within each quadrant of the CMA. The final virtual environment (n=12) used in the study were a result of focus groups and a pilot trial. Researchers in the focus groups identified the initial selection list (n=81). Then, researchers excluded virtual environments (n=38), which were perceived as neutral or received highly conflicting ratings where the focus groups could not agree on the emotion categorisation. Finally, for the remaining virtual environments (n=43), to ensure a diverse selection of the virtual environments, the research team voted for the most emotionally intense virtual environments among virtual environments that shared highly similar content. For example, only one virtual environment was selected among virtual environments containing baby animals such as puppies and kittens. In this process, 22 virtual environments were excluded. As a result, the pilot trial included twenty-one virtual environments.

Twelve volunteers (six females and six males) aged between 19 and 33 (M=24.17, SD=4.19) engaged in and rated the selected virtual environments (n=21). The volunteers watched the virtual environments in a randomised order. Volunteers rated each virtual environment using the following tools:

- Self-Assessment Manikin (SAM) (Bradley & Lang, 1994) is a well-established affective state measurement using picture-based of a cartoon-like manikin shapes (SAM) to plot basic the

CMA dimensions. The valence scale ranges from 1="sad" to 9="happy" pictures of SAM, while the arousal scale ranges from 1="calm" to 9="excited" pictures of SAM.

- The Visual Analog Scale (VAS) (Hawker et al., 2011) is a horizontal scale ranging across a continuum from 0 to 100, anchored by two verbal descriptors at each end. Using VAS, volunteers rated how they felt whilst engaging in virtual environments using a scale of: joy, happiness, calmness, relaxation, anger, disgust, fear, anxiousness and sadness.

The selected virtual environments are presented in Table 5.1, along with the participant's valence and arousal ratings.

Table 5.1: Selected Virtual Environments Arousal and Valence Ratings

CMA	Title	Rated Valence		Rated Arousal	
		M	SD	M	SD
High Arousal Positive	Rope Walking	6.42	1.38	7.17	1.34
	Brazilian Dancing	6.38	1.60	6.00	2.45
	Dancing with the Stars	6.67	1.51	6.00	1.26
Low Arousal Positive	Beautiful Resorts	7.83	1.59	3.83	2.55
	Pond in a Forest	7.08	1.44	3.58	2.68
	Cute Bunnies	7.42	1.68	3.50	2.58
High Arousal Negative	The Exorcist	3.75	2.18	6.75	1.86
	Alone in a Tent	3.83	2.21	6.50	2.39
	Zombies Eating Flesh	3.33	2.39	6.33	2.39
Low Arousal Negative	Post Terror Attacks	3.25	1.96	3.42	2.19
	Refugee Stories	2.75	1.76	3.50	2.02
	Refugee Boats	2.17	1.80	3.83	2.44

Figure 5.1: Examples of virtual environments used in the study

5.2.3 Data Collection

5.2.3.1 Participants

Thirty-four individuals (17 female and 17 male) aged between 18 and 61 years (M=25.0, SD=7.65) volunteered to take part in this study. 55.9% of participants (n=19) reported having used VR before, of which none have reported feeling motion sick amid or post-exposure to VR. On a Likert scale from one to seven on "How easily do you get motion or carsick?" where 1="never been motion sick" and 7="get motion sick very easily", participants reported they do not easily get motion sick (M=1.35, SD=1.12).

5.2.3.2 Apparatus and Setup

All computers and machines were set up in a "control room," while the "participant room" only contained a table and chair, a VR headset and headphones, a laptop, and ECG and GSR cables.

Headset & Headphones: The FOVE-0 (*Fove official website* 2021) headset and a set of headphones were wire-connected to a dedicated computer to stream the visual and auditory content. The FOVE-0 is a hands-free headset secured with its 3-point harness adjustable Velcro head straps. The headset has a WQHD OLED display (2560x1440 pixels) and renders at a frame rate of 60 fps with a field of view up to 100 degrees. The head orientation tracking system uses Inertial Measurement Units (IMU), and the eye tracking system uses infrared-based technology on each eye with tracking accuracy less than 1 degree at 120 fps and running at a sampling frequency of 60Hz. Headphones were used to stream auditory content.

Biopac System: The Biopac MP 150 system (*BIOPAC* 2020) was used to continuously acquire ECG (using the ECG100C hardware module) and GSR (using the EDA100C hardware module) signals and each were set to a unique channel. For both ECG and GSR, Biopac data acquisition software AcqKnowledge 4.1 (*ACQKNOWLEDGE* 2020) was used. The Biopac system was wire-connected to a dedicated computer. The participant's right arm was secured using Velcro tape to ensure minimal noise to ECG and GSR signals.

ECG: Continuous signals were acquired using a Lead II configuration (LEAD110-Series cables and 3-meter MEC110C extension cables) where electrodes were placed on the participant's right arm (Vin-) wrist and left calf (Vin+). Pre-gelled disposable electrodes (EL500-Series) were used, enhanced with electrode gel (GEL100) to increase the conductivity between the skin and the electrode and secured with an adhesive tape to ensure minimal noise.

GSR: Continuous signals were acquired using two disposable and adhesive electrodes (BIOPAC EL507) placed on the participant's right hands' index and middle fingers and wire-connected to the hardware module (using LEAD110A cables).

Lastly, an 11" Macbook Pro laptop and mouse were provided for the participant to fill all self-reported questionnaires digitally.

5.2.3.3 Physiological Measures

5.2.3.3.1 Eye-tracking

Eye tracking is the process of measuring either the point of gaze (where one is looking) or the motion of an eye relative to the head. Eye tracking data can be used to understand how people interact with visual content and has been used throughout literature in emotion recognition tasks (Lim et al., 2020). Eye-gaze metrics encompass various parameters, including blinking, fixation (a temporary pause of the eye during visual and cognitive processing, typically occurring over significant regions of interest), saccade (swift eye motion between fixations), and micro-saccade (small, jittery eye movements within a fixation) (Holmqvist & Andersson, 2017). Research indicates that eye tracking has potential as a robust method for evaluating emotional responses to stimuli. Eye tracking features (pupil size, pupil position and motion speed of the eye) have been used to detect four different types of emotion with an accuracy of 90% (Raudonis et al., 2013). Another study utilised pupillary and eye gaze behaviour measures in the analysis of emotional reactions to positive, negative, and neutral video clips, achieving an accuracy of 77.80% (Lu et al., 2015). In a survey carried out by Lim et al. (2020), it was found that a combination of eye-tracking features was required to obtain state-of-the-art results. They also found that the most used features were pupil diameter (Raudonis et al., 2013; Alhargan et al., 2017) and fixation duration (Tsang, 2016; Alhargan et al., 2017).

5.2.3.3.2 ECG

An ECG is a test that records the electrical activity of the heart. It measures the heart's electrical signals and has visible waves and spikes. The ECG waveform reflects the electrical activity of the heart and shows how well it is functioning. The ECG wave complex consists of different components, including the P wave, QRS complex, and T wave, each of which corresponds to a different stage of the heart's electrical cycle (Ashley & Niebauer, 2004). The ECG waveform is a valuable tool for monitoring heart health and can help healthcare providers make informed decisions about treatment. Heart rate variability (HRV) can be measured using information from the ECG components. HRV is a measure of the variation in time between successive heartbeats. HRV reflects the interplay between the sympathetic and parasympathetic nervous systems and provides information about the autonomic regulation of the heart. HRV has been used as an indicator of various health conditions, including stress, anxiety, depression, and heart disease. Higher HRV is generally considered a sign of good health, while lower HRV can indicate that the body is under stress or that there is an underlying health problem. ECG signals have been used effectively to detect emotion through literature (Katsigiannis & Ramzan, 2018; Abadi et al., 2015). For further discussion on the use of ECG to detect emotion, see sections 2.4 and 2.6.1.

5.2.3.3.3 GSR

GSR is a method of measuring the electrical conductance of the skin, which is an indicator of physiological arousal. It is commonly used as a measure of psychological or physiological arousal in response to various stimuli, such as emotional states, stress, or pain. The method works by applying a small electrical current to the skin and measuring the skin's resistance to the current, which changes as a result of sweat gland activity. GSR has been used effectively to detect emotion through literature (Koelstra et al., 2012; Udovičić et al., 2017). For further discussion on the use of GSR to detect emotion, see sections 2.4 and 2.6.1.

5.2.3.4 Procedure

Each participant undertook one 2-hour session in the laboratory. The following steps were then followed:

1. On arrival, verbal instructions protocol was given. This to protocol ensured that the instructions were consistent for all participants. Prior to the start of the session, participants were informed about the study but were not informed about the purpose or the hypotheses of the study. The consent form and the "participant profile & pre-exposure" questionnaire were then filled in.
2. ECG and GSR electrodes were then applied to the participant. Afterwards, the equipment was tested to ensure the electrodes picked up signals correctly, followed by a 3-minute rest recorded as a baseline. During rests (baseline or between virtual environments), participants were asked to close their eyes, breathe normally and try not to think of anything too exciting or stressful.
3. Next, participants were introduced to the VR headset; they were explained how the headset could be fitted and how to navigate virtual environments. Eye-gaze calibration was required using the standard FOVE-0 calibration program at this stage.
4. Participants were then asked to engage in the virtual environment from beginning to end.
5. At the end of each virtual environment, participants filled the "Post-Exposure" questionnaire then had a two-minute cool-down period (where they were asked to relax and sit quietly) before the next one.

Step 4 and 5 was repeated until participants engaged in all virtual environments. The order of the quadrants and the virtual environments within the quadrants were randomised. At the end of the session, participants were fully briefed about the study aims and received a £10 Amazon voucher.

5.2.4 Feature Extraction

The study included 34 participants engaging in twelve virtual environments, yielding a total of 408 trials. Data from eight participants were excluded from further analysis due to the poor quality in one or more of the data modalities recorded caused by technical problems. Therefore, the final dataset included 26 participants and 312 trials.

5.2.4.1 Eye Tracking Feature Extraction.

Eye tracking raw data per participant were generated for each virtual environment from beginning to end, including data for the left eye, right eye, and head rotation. The GazeParser library (Sogo, 2012) was used to extract eye-gaze features for analysis. In preparation for feature extraction, eye-gaze and head-movement data were used in conjunction to produce horizontal and vertical viewing angle (X, Y) data per eye. The feature extraction algorithm is based on the velocity threshold method or what is also known as velocity-based identification, where the algorithm distinguishes fixations (low velocity) from saccades (high velocity) based on gaze point-to-point velocities (Salvucci & Goldberg, 2000). Finally, statistical calculations were carried out for each feature. Considering that the virtual environments had different time lengths, the count (number of instances) were normalised by length (number of instances/time). Table 5.2 describes the eye tracking features that were extracted, brief description, threshold and the statistical calculations carried out for each feature, including Normalised Count (NormCount), Mean (M), Maximum (Max), Standard Deviation (SD) and Skewness (Skew); a measure of asymmetry in a distribution.

Table 5.2: List of Eye Tracking Features

Main Feature	Brief Description	Threshold	Statistical Metrics
Fixation	Temporal stillness in the eye movement over time	Fixation minimum duration=300ms	Number of fixations (NormCount) per second First fixation duration Duration (M, Max, SD, Skew)
Micro-Saccade	Intra-fixational movement where the eye jitters during a fixation	Micro-saccade minimum duration during a fixation=400ms	Number of micro-saccades (NormCount) per second Peak velocity (M, Max, SD, Skew) Direction (M, Max, SD, Skew) Horizontal amplitude (M, Max, SD, Skew) Vertical amplitude (M, Max, SD, Skew)
Saccade	Rapid motion of the eye from one fixation to another	Saccade velocity threshold=35ms Saccade acceleration threshold=400ms Saccade minimum duration=30ms Saccade minimum amplitude=5ms	Number of saccades (NormCount) per second Duration (M, Max, SD, Skew) Direction (M, Max, SD, Skew)
Blink	Eye closed	Blink minimum duration =50ms	Number of blinks (NormCount) per second Duration (M, Max, SD, Skew)

5.2.4.2 ECG Feature Extraction

Prior to feature extraction, the ECG signals were filtered using a high pass butter worth filter with a cut off filter frequency of 1hz and filter order of 2 to correct noise and baseline problems. Features are shown in the literature (Colomer Granero et al., 2016) to be reliable predictors of emotional state are used.

R-R intervals served as the base for the calculations many of our ECG features. R-R intervals refer to the time intervals between successive R waves of an ECG signal, used to measure heart rate variability. Examples of R peaks can be found in Figure 5.2. In order to find R-R intervals, peaks in data need to be identified and then before carrying out several steps of outlier detection and rejection. In order to detect peaks, a moving average is calculated using a window size of 0.75 seconds, peaks are then selected above the average. Beats per minute and the standard deviation of successive differences are then checked to see if they are in a normal range (40-200bpm), peaks are rejected if they fail this criteria. After these peaks had been selected, twelve time domain and features were selected using them. Along with these, the mean maximum and minimum signal amplitude were calculated and absolute power of three different power bands. Considering that the virtual environments had different time lengths, the count was normalised by length (number of instances/time) where relevant. A total of 18 features were extracted. Features were extracted using an implementation by heartpy (van Gent et al., 2019). All features and their descriptions can be found in table 5.3.

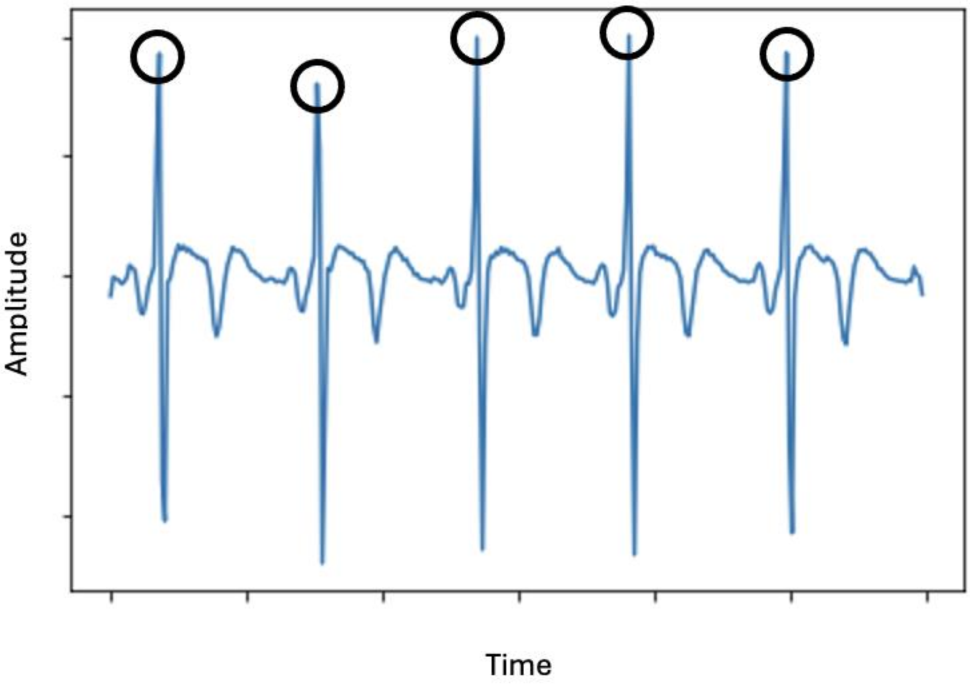


Figure 5.2: Example of an ECG signal and r peaks

Table 5.3: List of ECG Features

Feature	Brief Description
Signal Amplitude (M, Max, Min)	The voltage of electrical activity recorded
R-R Intervals (M, Med, Max, Min, SD, SDSD, RMSSD)	Time elapsed between two successive R-waves
ibi	Inter beat interval
bpm	Beats per minute
pnn50	Percentage of successive RR intervals that differ by more than 50ms
pnn20	Percentage of successive RR intervals that differ by more than 20ms
pnn50pnn20	Ratio of pnn50 and pnn20
VLF	Absolute power of the very-low-frequency band (0.0033–0.04 Hz)
LF	Absolute power of the low-frequency band (0.04–0.15 Hz)
HF	Absolute power of the high-frequency band (0.15–0.4 Hz)

5.2.4.3 GSR Feature Extraction.

A low pass butter worth filter that removes muscle noise which allows the detection of sweating peaks to be more accurate, was first applied. A cutoff filter frequency of 1hz and a 2nd order filter. Features are shown in the literature (Colomer Granero et al., 2016) to be reliable predictors of emotional state are used. Considering that the virtual environments had different time lengths, the count was normalised by length (number of instances/time) where relevant. A total of eight features were extracted, as described in table 5.4.

Table 5.4: List of GSR Features

Feature	Brief Description
Skin Conductance level (M, Max, Min, SD, Var)	Tonic level of electrical conductivity of skin
Number of local minima per second	Number of valleys per second
Number of local maxima per second	Number of peaks per second
Peaks/Time	The ratio of peaks and time

5.2.5 Model Development

5.2.5.1 Pilot

The pilot process and techniques utilised in this study were the same as those described in Chapter 4. The reason for this selection was based on the similarities in the data being dealt with. Although larger than that used in Chapter 4, the dataset is still relatively small and contains minimal features.

Similar to Chapter 4, performance metrics, interpretability, computational requirements, model complexity, feature importance methods and support from the literature were considered when selecting the model to proceed with. With this research, a higher importance was put on pilot performance and less on interpretability compared to chapter three. This shift is driven by research objective I, which aims to validate the dataset and assess the effectiveness of VR as an emotional stimulus. Our focus on these objectives led us to favour techniques that performed well in the pilot trial and techniques employed in comparable studies, enhancing the quality of our comparisons and contributing to a more robust assessment.

The results of this pilot study can be found in Table 5.5 and Figure 5.3. K-Nearest Neighbour (KNN) and Decision Trees (DT) are clearly outperformed by Logistic Regression (LR), Support Vector Machines (SVM) and Random Forest (RF) and were therefore not in consideration to progress with. RF outperforms both SVM and LR, but SVM has a smaller standard deviation, suggesting it could be a more reliable algorithm moving forward when generalising to unseen data and various validation methodologies.

Interpretability of the model was not a major factor in deciding which algorithm to use, because these results were mainly used to demonstrate the effectiveness of immersive stimuli against other non-immersive stimuli and not as a potential exploration of an algorithm to use in clinical settings. LR and RF are more interpretable than SVM. For these reasons, computational requirements and model complexity did not also have a major impact in the decision process. RF and LR are both less computationally expensive than SVM.

Overall, SVM was selected mainly because it produced good results in the pilot trial. In addition to this, SVM will also allow for a consistent comparison to established benchmarks. We wanted to compare our dataset to four other well-known emotional datasets that provide baseline ML results, DEAP (Koelstra et al., 2012), DREAMER (Katsigiannis & Ramzan, 2018), DECAF (Abadi et al., 2015) and MAHNOB-HCI (Soleymani et al., 2012). The results reported by DEAP were achieved using a naïve bayes classifier. The results reported by the other three, including DECAF which achieved the highest accuracy, all used SVMs. Therefore, the use of an SVM in our study would provide a good comparison.

Table 5.5: Pilot Results

Model	Mean Accuracy	Standard Deviation
Logistic Regression (LR)	0.75	0.10
Random Forest (RF)	0.78	0.11
Decision Tree (DT)	0.63	0.07
Support Vector Machine (SVM)	0.74	0.05
K Nearest Neighbour (KNN)	0.60	0.08

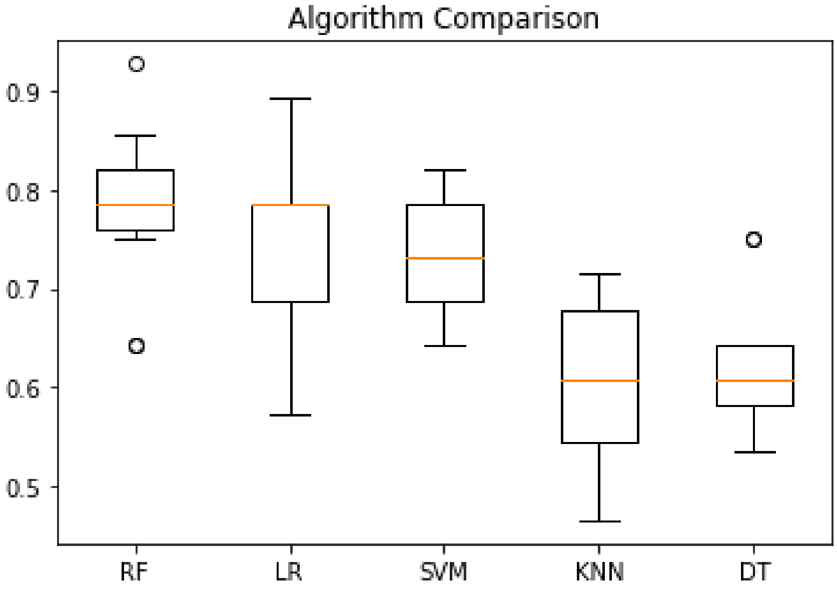


Figure 5.3: VREED Pilot Results

5.2.5.2 Model Evaluation

Three different validation methods were used on this data, a holdout set, ten-fold nested cross validation and leave one participant out nested cross validation. For the holdout set, ten-fold cross-validation was performed on 90% of the data to train and validate our selected model, leaving 10% for unseen testing. For ten-fold cross nested cross validation, the complete dataset was randomly split, meaning that data from the same participant may appear in both the training and validation sets. The final validation method, leave on participant out nested cross validation ensured data from a single participant was not in both the training and validation data set.

Hyper parameters were tuned in the nested cross validation using randomised grid search. the list of hyperparameters and values considered can be found in table 5.6.

Overall, we looked at three classification tasks:

i) Four class classification of the four CMA quadrants

ii) Binary high and low arousal

iii) Binary high and low valence

Table 5.6: List of SVM Parameters and Values

Parameter	Values	Description
C	[0.1,1, 10, 100]	a regularisation parameter that determines the cost of misclassifying training examples.
Kernel	['rbf', 'poly', 'sigmoid']	a function that transforms the input data into a higher-dimensional feature space
Gamma	[1,0.1,0.01,0.001]	determines the influence of each training example on the decision boundary

5.2.5.3 Feature Importance

We used the same technique to obtain feature importance as in Chapter 4. The other technique that could be applicable to get the feature importance is permutation importance. It is a model-agnostic method that works by randomly permuting (randomly changed order or position of element) the values of a feature in the test data and measuring the decrease in the model's performance as a result of the permutation. The larger the decrease in performance, the more important the feature is considered to be.

One advantage of permutation importance is that it directly measures how much a feature contributes to a model's performance, while forest-based feature importance only provides a relative ranking of feature importance within the context of a random forest model.

However, a benefit of forest-based importance is that it can capture non-linear relationships between features and the target variable, while permutation importance can only capture linear relationships. When a feature is randomly permuted, the model assumes that the change in the feature's value is equivalent to the change in the feature's importance. This assumption only holds if the relationship between the feature and the target variable is linear. If the relationship is non-linear, the permutation of the feature may not result in a significant decrease in performance, even if the feature is important to the model. Also, permutation importance can cause confusing results if the model contains correlated features. When two features are highly correlated, permuting one feature may not affect the model's performance as much as expected because the other correlated feature can still capture the same information. This can lead to an underestimation of the importance of the permuted feature. On the other hand, permuting a correlated feature may also have a larger impact on the model's performance if the other correlated feature is not as important. This can lead to an overestimation of the importance of the permuted feature.

Tree-based feature importance measures can handle correlated features better than permutation importance because they use DTs to calculate feature importance. DTs can capture the interactions between features and model non-linear relationships between features and the target variable. Therefore, DT-based models like RF can identify important features even if they are highly correlated with other features in the dataset. This method doesn’t provide direct insights into the inner workings of the SVM model. However, it can give us more meaningful insights into the data.

5.3 Results

5.3.1 Validation of Dataset

This analysis aimed to understand whether virtual environments triggered the desired arousal and valence effects among the participants. Table 5.7 describes the valence and arousal ratings using SAM in each quadrant. For the valence dimension, both negative and positive virtual environments were perceived as intended (low arousal negative; M=2.46, high arousal negative; M=4.29, low arousal positive; M=6.48, high arousal positive; M=6.46). For the arousal dimension, the intended low-arousal virtual environments were perceived as intended (low arousal negative; M=3.52, low arousal positive; M=2.51). However, for the intended high arousal virtual environments, the intended high arousal negative virtual environments were perceived as intended but the high arousing positive was not (high arousal negative; M=6.00, high arousal positive; M=3.80).

Table 5.7: Ratings of Arousal and Valence per CMA Quadrant

Intended CMA	**Rated Arousal**		**Rated Valence**	
Quadrant	M	SD	M	SD
High Arousal Positive	3.80	1.97	6.46	0.77
Low Arousal Positive	2.51	1.12	6.48	1.00
High Arousal Negative	6.00	1.60	4.29	1.59
Low Arousal Negative	3.52	1.69	2.46	1.06

Two-way repeated-measures ANOVA, followed by Tukey's HSD test were carried out to determine the significance of engaging in virtual environments in the four CMA quadrants over both, the valence and arousal dimensions.

The rated valence significantly differed in the four quadrants of the CMA, F(132, 2)=21.62, p<.001. Tukey's HSD test indicated that the mean of rated valence in negative virtual environments (M=3.38, SD=1.63) was significantly lower than positive virtual environments (M=6.47, SD=0.89), t(132, 2)=-15.59, p<.001. The mean of rated valence in low arousal virtual environments (M=4.44, SD=2.71) was significantly lower than high arousal virtual environments (M=5.36, SD=1.66), t(132, 2)=-4.55, p<.001; meaning that participants rated high arousal virtual environments significantly more positively than low arousal virtual environments.

The rated arousal significantly differed in the four quadrants of the CMA, F(132, 2)=4.55, p=.035. Tukey's HSD test indicated that the mean of rated arousal in low arousal virtual environments (M=3.02, SD=1.51) was significantly lower than high arousal virtual environments (M=4.92, SD=2.09), t(132, 2)=-6.71, p<.001. The mean value of rated arousal in negative virtual environments (M=4.76, SD=2.05) was significantly higher than positive virtual environments (M=3.16, SD=1.72), t(132, 2)=5.71, p<.001; meaning that negative virtual environments were perceived as significantly more arousing than positive virtual environments.

To sum up, participants encountered four clearly defined emotional states along the arousal spectrum (high, low) and valence spectrum (negative, positive). Despite perceiving high arousal positive virtual environments as less arousing, their arousal ratings in this category were notably higher than those for low arousal positive virtual environments.

5.3.2 Machine Learning Results

Three classification tasks, two binary and a four-class problem, were carried out: high/low arousal (binary), positive/negative valence (binary) and the four classes in accordance with the four quadrants of the CMA. Based on initial tests with a range of classifiers and drawing from literature (Domínguez-Jiménez et al., 2020; Katsigiannis & Ramzan, 2018), SVM was selected and was used for the analysis and three different validation methods were used.

Results for all validation methods can be found in tables 5.8, 5.9 and 5.10. Nested 10-fold cross validation produced the best results but all results were comparable. The model may generalise effectively even when validation methods vary could be due to the dataset exhibiting homogeneity in participant responses or emotional patterns being relatively stable across the VR experiences. Such consistency of results across validation techniques suggests a degree of robustness in the model's predictive capability.

The results from all validation methods showed the benefits of using combined modalities especially in the four-class classification task. The eye tracking features substantially outperforms the ECG features for arousal and four-class classification but performs similarly in recognising valence. Eye Tracking features outperform the GSR features in valence and four-class classification tasks and they perform equally in the arousal classification task. The combination of features of all three modalities achieves the highest accuracy than all of the individual modalities for all three classification tasks.

Tables 5.8, 5.9 and 5.10 present a summary of these results, specifically highlighting the results obtained using similar modalities as VREED as well as the modalities which achieved the best accuracy for each dataset. Direct comparisons with the literature work must be proceeded with caution due to the differences in experimental setups, apparatus, conditions and feature extraction and analysis methodologies. Furthermore, research in affective computing examining the use of VR is scarce. However, some similarities with studies which examined affect in non-immersive mediums can be observed. The baseline results achieved with the VREED dataset are consistent with other baseline results reported in other affective datasets that used non-immersive mediums. Our results are comparable with DEAP (Koelstra et al., 2012) where EEG and peripheral physiological responses of which included GSR, blood volume pressure, respiration pattern, skin temperature and Electromyography (EMG) were recorded whilst engaging in video clips using a PC monitor. 104 features were extracted from these modalities. Using peripheral features (excluding EEG), the authors obtained accuracy of 57% when classifying arousal and 62.7% accuracy when classifying valence which compare to our 87.5% and 56.25% respectively using only GSR features. When these features were fused with EEG and multimedia content analysis features, higher accuracy was obtained (61.6% for arousal & 64.7% for valence). Dreamer (Katsigiannis & Ramzan, 2018) contains ECG and EEG data, the classification accuracy using the ECG data with SVM reported on the baseline analysis was 62.37% for both arousal valence compared to our 75% and 78% accuracy for arousal and valence. DECAF (Abadi et al., 2015) includes various peripheral physiological responses including ECG. Using late fusion, arousal was classified with 73.5% and valence with 77.5% accuracy. The MAHNOB-HCI dataset (Soleymani et al., 2012) includes four peripheral nervous system signals: GSR, respiration amplitude, skin temperature, and ECG along with EEG and eye-gaze data. A fusion of modalities performed best followed by the eye-gaze data, achieving 68.8% and 63.5% for binary classification of arousal and valence. Our eye tracking results compare well achieving 87.5% and 78.13% for arousal and valence. In conclusion, our baseline results compare well to other well-established emotion recognition datasets, proving VREED to be a robust dataset for further research ML analysis.

The dataset used in this study has several limitations that should be considered when interpreting the results. Firstly, the dataset was relatively small, which may increase the risk of overfitting. Secondly, the participants were drawn from a relatively small geographical and social area, which may not capture the diversity of emotional responses across broader populations. This lack of diversity could affect the generalisability of the findings, as the model may not account for cultural, demographic, or contextual variations in emotional experiences. Additionally, the focus on a specific set of VR experiences may limit the applicability of the findings to other types of virtual environments (such as more interactive games) or emotional contexts.

Table 5.8: ML Results Using Different Modalities Holdout Methodology

Modality	Arousal	Valence	4-Class
ECG	75%	78%	50%
GSR	87.5%	56.25%	53.13%
Eye Tracking	87.5%	78.13%	62.5%
ECG + GSR + Eye Tracking	90.63%	84.38%	71.88%

Table 5.9: ML Results Using Different Modalities Nested 10 fold cross validation Methodology

Modality	Arousal	Valence	4-Class
ECG	73% (SD=6%)	68% (SD=6%)	45% (SD=9%)
GSR	89% (SD=5%)	65% (SD=7%)	62% (SD=7%)
Eye Tracking	85% (SD=6%)	75% (SD=7%)	66% (SD=12%)
ECG + GSR + Eye Tracking	92% (SD=8%)	78% (SD=9%)	73% (SD=6%)

Table 5.10: ML Results Using Different Modalities Nested Leave-One-Participant-Out Methodology

Modality	Arousal	Valence	4-Class
ECG	69% (SD=17%)	63% (SD=14%)	41% (SD=19%)
GSR	88% (SD=10%)	64% (SD=10%)	61% (SD=15%)
Eye Tracking	83% (SD=13%)	75% (SD=15%)	64% (SD=15%)
ECG + GSR + Eye Tracking	91% (SD=12%)	78% (SD=14%)	72% (SD=15%)

Table 5.11: Precision, Recall and F1 Scores for all Modalities Holdout Methodology

Modality	Arousal			Valence			4-Class		
	P	R	F1	P	R	F1	P	R	F1
ECG	0.75	0.75	0.75	0.79	0.78	0.78	0.52	0.5	0.5
GSR	0.88	0.88	0.87	0.62	0.56	0.53	0.53	0.53	0.52
Eye Tracking	0.88	0.88	0.87	0.74	0.74	0.72	0.63	0.62	0.62
ECG + GSR + Eye Tracking	0.91	0.91	0.91	0.86	0.84	0.84	0.76	0.72	0.72

Table 5.12: Precision, Recall and F1 Scores for all Modalities Nested 10 fold cross validation Methodology

Modality	Arousal			Valence			4-Class		
	P	R	F1	P	R	F1	P	R	F1
ECG	0.72 (0.06)	0.74 (0.09)	0.73 (0.06)	0.69 (0.06)	0.69 (0.05)	0.69 (0.05)	0.45 (0.09)	0.44 (0.10)	0.45 (0.09)
GSR	0.85 (0.05)	0.89 (0.06)	0.88 (0.05)	0.65 (0.07)	0.65 (0.07)	0.65 (0.07)	0.62 (0.07)	0.63 (0.05)	0.62 (0.05)
Eye Tracking	0.87 (0.05)	0.82 (0.07)	0.85 (0.06)	0.72 (0.07)	0.73 (0.06)	0.072 (0.07)	0.68 (0.10)	0.65 (0.08)	0.66 (0.06)
ECG + GSR + Eye Tracking	0.91 (0.09)	0.91 (0.08)	0.91 (0.08)	0.78 (0.06)	0.77 (0.05)	0.78 (0.05)	0.74 (0.06)	0.74 (0.06)	0.74 (0.06)

Table 5.13: Precision, Recall and F1 Scores for all Modalities Nested Leave-One-Participant-Out Methodology

Modality	Arousal			Valence			4-Class		
	P	R	F1	P	R	F1	P	R	F1
ECG	0.70 (0.17)	0.69 (0.17)	0.69 (0.16)	0.64 (0.15)	0.63 (0.14)	0.63 (0.14)	0.41 (0.20)	0.41 (0.19)	0.41 (0.19)
GSR	0.89 (0.10)	0.88 (0.10)	0.88 (0.10)	0.69 (0.13)	0.64 (0.10)	0.64 (0.12)	0.61 (0.17)	0.61 (0.15)	0.61 (0.15)
Eye Tracking	0.85 (0.13)	0.83 (0.13)	0.84 (0.13)	0.76 (0.15)	0.75 (0.15)	0.75 (0.15)	0.65 (0.17)	0.64 (0.15)	0.64 (0.14)
ECG + GSR + Eye Tracking	0.92 (0.11)	0.91 (0.11)	0.91 (0.11)	0.80 (0.15)	0.78 (0.14)	0.78 (0.14)	0.74 (0.15)	0.72 (0.15)	0.72 (0.16)

Table 5.14: Results Comparison to Literature

Dataset	Modality	Arousal	Valence
VREED	ECG	75%	78%
	GSR	87.5%	56.25%
	Eye Tracking	87.5%	78.13%
	ECG + GSR + Eye Tracking	**90.63%**	**84.38%**
DEAP	Peripheral	57%	62.7%
	MCA	65.1%	61.8%
DREAMER	ECG	62.37%	62.37%
	ECG + EEG	62.32%	61.84%
DECAF	Peripheral	55%	60%
	Late Fusion	73.5%	77.5%
MAHNOB-HCI	Peripheral	46.2%	45.5%
	Eye Gaze	63.5%	68.8%
	Fusion	67.7%	76.1%

5.3.3 Feature Importance Results

5.3.3.1 ECG

The autonomic nervous system (ANS) plays a crucial role in regulating arousal and valence, and the nerve endings of the ANS within the cardiac muscle play a major role in the cardiac output because they affect the rhythm at which the muscle pumps blood (Agrafioti et al., 2012). The ANS has two systems: the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). The SNS is responsible for the body's "fight or flight" response, which is activated in response to perceived threats or danger (Sympathetic nervous system (SNS): What it is & function 2023). This response leads to increased heart rate, rapid breathing, and heightened arousal. The PNS, on the other hand, is responsible for the body's "rest and digest" response, promoting relaxation and decreasing arousal (Parasympathetic Nervous System (PSNS): What it is & function 2023). When the sympathetic system dominates the parasympathetic in times of stress, resulting in the following reactivity effects (Bhuiyan et al., 2008):

- *Automaticity*: the spontaneous generation of electrical impulses by pacemaker cells in the heart, leading to an increased heart rate.

- *Contractility*: the heart muscle's ability to contract or shorten in response to an electrical impulse generated by the pacemaker cells. The force of contraction is increased during times of stress. This results in an increased amplitude of the ECG signal.

- *Conduction Rate*: also known as heart rate, is the number of times the heart contracts or beats per minute increases.

- *Excitability*: increased perceptiveness to internal and external stimuli, possibly leading to ectopic beats that would be captured in HRV features

These effects are captured in our features. Our feature importance results show that the signal amplitude is the most important feature, highlighting that the contractility provides the most information to the model and warrants focus on the further development of features. Here we have demonstrated that our ECG features and feature importance’s align with features reported throughout medical literature to correlate with emotion and its components (arousal and valence). We have also presented possible medical reasons for the importance’s of these features.

Table 5.15: ECG Feature Importance Rankings

Feature	Importance
Max Signal	0.064
Min Signal	0.054
LF	0.05
Min R-R Interval	0.05
Inter Beat Interval	0.05
HF	0.049
BPM	0.049
Max R-R Interval	0.049
Median R-R Interval	0.049
Mean R-R Interval	0.048

5.3.3.2 GSR

Reticular activation, also known as the reticular activating system (RAS), is a network of neurons located in the brainstem that plays a crucial role in regulating wakefulness, arousal, and attention. GSR is a good indicator of reticular activation and is, therefore, linked to the energetic dimension of emotion. Research has shown that the amplitude of electrodermal responses increased linearly with arousal, regardless of valence (Sequeira et al., 2009). This can be seen in our research as GSR is a major contributor to the arousal model. Shukla et al (2021) also report that amplitude is the most significant feature for the recognition of both arousal and valence. This contrasts with our results; although the amplitude still has an important role, the top three of our features relate to the number of peaks and valleys, not their amplitude. This could be due to VR being used as stimuli, creating different types of increased arousal compared to non-immersive stimuli; further research needs to be carried out to investigate this.

Table 5.16: GSR Feature Importance Rankings

Feature	Importance
Ratio of Peaks to Time	0.27
Number of Peaks	0.13
Number of Valleys	0.12
Average	0.11
Minimum	0.1
Maximum	0.1
Variance	0.09
Standard Deviation	0.09

5.3.3.3 Eye Tracking

Human eye movement behaviours could reflect human emotional states (Lance & Marsella, 2008). Lim et al (2020) produced a survey that reported that no clear eye-tracking feature or a combination of features is most beneficial for emotion recognition tasks. However, when this survey was produced, no research reported detecting specific emotions in VR using eye-tracking technology. It was stated that compared to non-immersive stimuli, VR would arguably provide a more vivid experience. We have found that using VR stimuli, micro saccade, a small, involuntary eye movement that occurs during visual fixation, plays the biggest role in detecting emotion. This could be due to more natural eye movements being produced in VR compared to non-immersive stimuli as the participant is surrounded by the stimuli and can’t be distracted by the surrounding scene.

Table 5.17: Eye Tracking Feature Importance Rankings

Feature	**Importance**
Number of Micro Saccade	0.038
Standard Deviation Micro Saccade Peak Velocity	0.033
Max Micro Saccade Direction	0.031
Max Micro Saccade Vertical Amplitude	0.25
Max Saccade Duration	0.25
Max Saccade Length	0.24
Max Saccade Direction	0.24
Skew Micro Saccade Vertical Amplitude	0.24
First Fixation Duration	0.24
Standard Deviation Blink Duration	0.235

5.3.3.3 Combined Modalities

Looking at the combined feature importance allows us to further understand how the features interact and which modalities and features have the most impact overall for both arousal and valence binary classification and the four class classification tasks.

When looking at the interaction of features across modalities, GSR features seem to benefit highly from the addition of the ECG and eye-tracking features. For all three classification tasks, GSR features are prevalent throughout each top ten but do not provide the best classification results when using individual modalities to train ML models. Standard Deviation Micro Saccade Vertical Amplitude is the fifth most important feature in the four-class and arousal classification and the fourth most important feature in valence classification but does not appear in the top ten features in the single eye tracking modality classification feature importance’s again suggesting that this feature benefits from contextual information from ECG and GSR features. It is important to note that there is no ECG feature within the top ten features for four-class classification; the highest-ranked feature is maximum amplitude which ranks twenty-eighth.

Table 5.18: Combined Modalities 4 Class Feature Importance Rankings

Feature	Importance
Ratio of Peaks to Time (GSR)	0.071
Number of Peaks (GSR)	0.028
Number of Micro Saccade (Eye)	0.028
Number of Valleys (GSR)	0.025
Standard Deviation Micro Saccade Peak Velocity (Eye)	0.022
Average (GSR)	0.021
Max Micro Saccade Direction (Eye)	0.018
Standard Deviation Micro Saccade Vertical Amplitude* (Eye)	0.017
Max Micro Saccade Vertical Amplitude (Eye)	0.017
Maximum (GSR)	0.016

Table 5.19: Combined Modalities Arousal Feature Importance Rankings

Feature	Importance
Ratio of Peaks to Time (GSR)	0.130
Number of Peaks (GSR)	0.050
Number of Valleys (GSR)	0.045
Average (GSR)	0.033
Standard Deviation Micro Saccade Peak Velocity (Eye)	0.021
Minimum (GSR)	0.021
Max Micro Saccade Direction (Eye)	0.019
Max Saccade Direction (Eye)	0.019
HF (ECG)	0.017
Maximum (GSR)	0.017

Table 5.20: Combined Modalities Valence Feature Importance Rankings

Feature	Importance
Ratio of Peaks to Time (GSR)	0.037
Number of Micro Saccade (Eye)	0.032
Standard Deviation of Saccade Direction (Eye)	0.025
Standard Deviation Micro Saccade Peak Velocity (Eye)	0.024
Skew Micro Saccade Peak Velocity (Eye)	0.019
BPM (ECG)	0.019
IBI (ECG)	0.018
Max Micro Saccade Direction (Eye)	0.016
Standard Deviation Micro Saccade Vertical Amplitude (Eye)	0.016
MeanRR (ECG)	0.016

5.4 Discussion

5.4.2 Research Question I: Understand whether VR is a Reliable Stimulus for Emotion Recognition.

We have demonstrated the reliability of VR as a stimulus for emotion recognition via questionnaire data and statistical analysis. Analysis of the questionnaire showed that participants reported that they felt the correct emotion for each virtual experience (see 5.3.1). In addition, our baseline ML results outperformed baseline results published by datasets that utilised non-immersive stimuli (see 5.3.2).

5.4.3 Research Question II: Investigate whether Robust ML Models can be Developed to Recognise Emotion within Virtual Environments and how they Compare to ML Models Built Using Non-immersive Stimuli.

There are many publicly available affective datasets (Abadi et al., 2015; Katsigiannis & Ramzan, 2018; Koelstra et al., 2012; Schmidt et al., 2018; Soleymani et al., 2012). These datasets are not directly comparable (due to different sensor modalities and methodologies), but the ML results obtained using VREED are promising and show an improvement in accuracy in similar classification problems. We achieved 90.63%, 84.38% and 71.88% accuracy (when utilising holdout set validation) in classifying arousal, valence and the four sections of the CMA quadrant, respectively. The highest classification results achieved on baseline results produced by non-immersive stimuli are 73.5% and 75.5% accuracy when classifying arousal and valence (see more in 5.3.2).

5.4.4 Research Question III: Investigate the Specific Features of Data that Indicate Emotion and Compare these to Affective Computing and Medical Literature to make the ML Models More Transparent.

In our feature importance analysis, we have reported the feature importance of the ECG, GSR, eye tracking, and the combination of all modalities (see 5.3.3). We have compared these importance's to medical and affective computing literature. Generally, the feature importance highlighted features that are also used consistently throughout the literature in both fields. This gives further insights into the data.

Chapter 6: Real-Time Emotion Detection During Virtual Reality Experiences and XAI for Interpretation of Raw Time Series Physiological Signals

6.1 Introduction

Deep learning is a subfield of machine learning (ML) that is showing improved results and holds some benefits over traditional ML techniques. One of the key benefits of deep learning is that it can automatically learn useful features and representations from raw data, without the need for manual feature engineering. This can be especially useful for tasks such as image and speech recognition, where manually designing features can be difficult or time-consuming. Deep learning models are highly flexible and can be trained on a wide variety of tasks, including classification, regression, and generation. These benefits lend itself to analysing sets of raw time-series data.

The aim of this research is to use smaller chunks of time series physiological data in order to provide real-time detection. Real-time detection will become important, especially in the field of human-computer interaction. As the interaction between humans and computers becomes increasingly common, the development of emotional intelligent systems is becoming increasingly important (Jaiswal & Nandi, 2019). From a healthcare perspective, it is important to be able to measure emotional response during the course of psychotherapy (Sloan & Kring, 2007) and therefore, real-time detection will contribute towards future virtual therapy where a clinician may not be present. Further discussion on the role that emotional response plays during therapy can be found in section 2.4 (Emotion and engagement for mental health therapies).

Deep learning models and their ability to learn highly complex representations comes with drawbacks, the main criticism of deep learning models is that they are a black box. This affects how they can be used in the real world and their user's acceptability and trust in them. With this research, the aim is to start to describe and gain insight into this black box.

The application of explainable AI (XAI) methods for emotion recognition has primarily focused on modalities such as speech, text, and video (Mehta & Passi, 2022; Khalane et al., 2023; Heimerl et al., 2022). These studies consistently find that XAI techniques can help improve model transparency and interpretability for emotion classification tasks for these data types. However, few studies have examined the use of XAI for emotion recognition from multimodal physiological signals. Research into XAI for physiological data in other domains demonstrates its potential for providing insights into how the AI model makes predictions based on physiological data, even if such approaches are still emerging (Loh et al., 2023; Gouverneur et al., 2023). However, XAI has not yet been extensively explored for deep learning-based emotion recognition using physiological signals such as ECG, GSR, and eye tracking, particularly in the context of VR stimuli. As such, our work aims to develop XAI methods to explain predictions and learned representations of deep neural networks for real-time emotion classification from physiological signals. See section 2.6 for further discussion on XAI in emotion recognition systems.

In this research, we aim to develop a deep learning model that can detect emotion in real-time by building on top of research carried out initially in the previous chapter. The previous chapter utilises data from VREED, an emotion dataset that includes electrocardiogram (ECG), galvanic skin response (GSR) and eye-tracking signals. A baseline ML analysis was carried out, and the results were promising. We aim to expand on this by developing a deep learning model that uses short segments of data to detect emotion in real time. The results of our model and the baseline analysis will be compared. We will also present an XAI methodology to provide an explanation and more understanding of the predictions our model gives using global and local explanations. Our research objectives were as follows:

i) Investigate if the real-time emotions of VR users can be detected using raw time series physiological signal data.
ii) Investigate how can XAI be used to provide explanations for these predictions.
iii) Compare the XAI outputs to the feature importance of the baseline model and medical literature to further understand the inner workings of the model.

6.2 Methodology

6.2.1 Data Preparation

6.2.1.1 Dataset

VREED (VR Eyes: Emotions Dataset) (Tabaa et al., 2021) has been used for this analysis. Please refer to Chapter 4 for more information on this dataset. We will compare a deep learning approach to the preliminary ML analysis presented Chapter 4.

6.2.1.2 Pre-processing

6.2.1.2.1 Sliding Windows

The data from VREED was processed to make it suitable for deep learning and real-time detection. The data was down-sampled to 60hz and separated into 5-second sliding windows with a step of 1 second; research suggests that a 3-6 second window works and could be optimal for emotion recognition tasks using physiological signals (Ayata et al., 2016; Cai et al., 2021). Research has also shown that window sizes as small as 1-second of physiological data can be used successfully in emotion recognition tasks (Granato et al., 2018). VREEDS stimulus was selected by experts and the overall, colours, sounds and theme of the videos in VREED should induce the correct emotion from the beginning (Udovičić et al., 2017). We will also be testing a window size of 2.5 and 10 seconds to understand how different size windows compare to the 5-second window.

6.2.1.2.2 Separating Data for Validation and Testing

To develop and test our model, we used stratified sampling: In this approach, the data is split into training, validation, and test sets.

Three random participants were chosen to be separated (5356 total data points) for unseen testing, these participants were not used for any of the training and parameter tuning, therefore testing the generalisability of the model on unseen participants. The training dataset consisted of 37,471 data points. For model development, we split a random 5% of the data for validation on each epoch, this data contained data from multiple participants. A breakdown of these splits can be found in Figure 6.1.

Only using 3 participants also allows a focused approach to the analysis and XAI of the predictions. It is common for XAI proof of concepts to be carried out on minimal examples to allow in-depth analysis of local explanations; for example, La Ferla (2023) investigated a CNN that predicted breast cancer. In the analysis, 20 mammograms were used to provide a proof of concept to begin discussions on the implementation of XAI within the clinical settings. 100 examples from each participant (300 total) new looked at in this analysis.

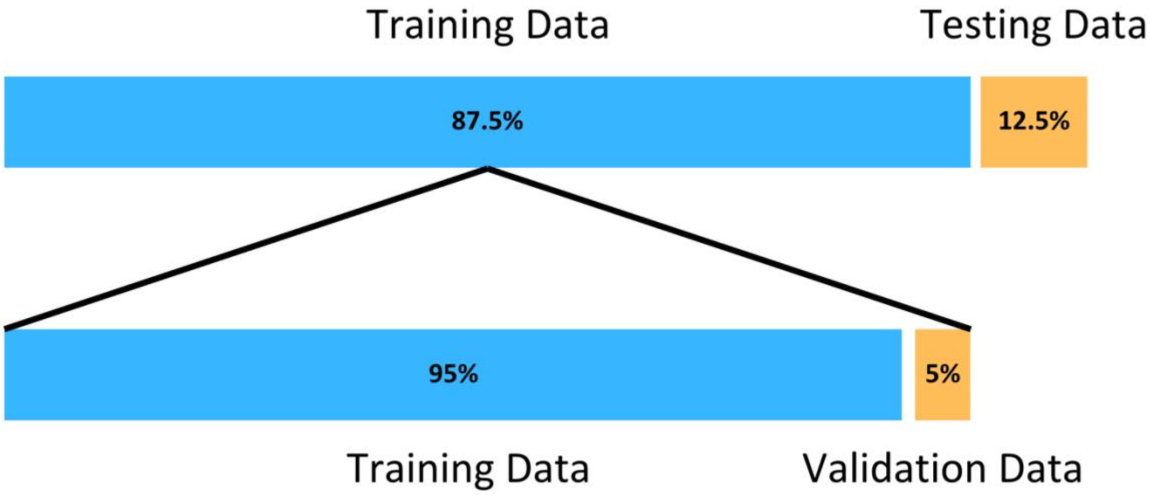


Figure 6.1: Train Test Spilt

6.2.2 Model Development

Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are both types of deep learning algorithms that have been applied to a variety of tasks involving time series data. In general, deep learning techniques such as CNNs and RNNs have been found to outperform traditional ML techniques on tasks involving time series data (Ismail Fawaz et al., 2019; Siami-Namini et al., 2018). This is partly because deep learning models can automatically learn features and representations from the data, rather than relying on manually engineered features. As a result, deep learning models can achieve state-of-the-art results on a wide range of time series tasks. For these reasons, both CNNs and RNNs were developed in this study.

During development, we looked at CNNs and LSTMs individually and analysed their strengths and weaknesses before considering a hybrid approach. We are aware that there are various deep learning models however CNN, LSTM, and CNN+LSTM models are the most commonly used (Rim et al., 2020).

6.2.2.1 CNN Models

A CNN is a type of artificial neural network that lends itself to processing data that has a grid-like topology, such as an image. CNNs are composed of multiple layers of interconnected nodes, with each layer learning to recognize patterns and features in the data.

Overall, CNNs have proven to be highly effective for tasks such as image classification, object detection, and segmentation, and have achieved state-of-the-art results in these tasks (Li et al., 2022). CNNs have also been used effectively for time series classification, since they are highly noise-resistant models, and are able to extract informative, deep features.

CNNs are used frequently when analysing ECG signals for many different tasks, one of the most studied tasks is detecting arrhythmias. CNNs are one of the most popular deep-learning architectures for this (Dong et al., 2022; Ebrahimi et al., 2020). Arrhythmia refers to an abnormal rhythm of the heart, which can occur when the electrical impulses that regulate the heartbeat are disrupted or irregular. A reason deep learning architectures are used in these classification tasks is that due to their capabilities in preserving temporal variation of the signal. This would be important in the real-time classification of emotions to understand emotion and signal change over time. As regards to emotion recognition, a self-supervised CNN has been developed (Sarkar & Etemad, 2022) that only utilises ECG signals that outperform baseline results for DREAMER (Katsigiannis & Ramzan, 2018) among other well-known emotional datasets. For DREAMER, a self-supervised CNN achieved 85.9% and 85% accuracy when classifying arousal and valence, respectively; the original SVM achieved 62.4% for both arousal and valence Classification.

CNNs have also been used that only utilise GSR signals that outperform “shallow” ML methods on the AMIGOS dataset (Santamaria-Granados et al., 2019). This method achieved 71% and 75% accuracy for arousal and valence classification; the closest shallow method for arousal classification was an SVM, which achieved 69% accuracy; and for valence classification, the closest “shallow” method was a K-nearest neighbour (KNN) which achieved 69% accuracy.

CNNs have also been used successfully with eye-tracking data. An adjusted LeNet-5 architecture (Lecun et al., 1998) which is a relatively simple architecture that consists of 7 total layers and 3 convolutional layers has been used to classify two different web interfaces for browsing news data (Yin et al., 2018). The previous research mentioned was not related to emotion recognition but a fully connected neural network has been used to classify excitement (arousal) in real-time (Abdessalem et al., 2019) It is reasonable to believe that results obtained from a fully connected network could be improved using more advanced architectures and that a CNN could achieve good emotion classification results.

6.2.2.1.1 CNN Development

The overall strategy for model development was to start relatively simple with an adjusted lenet-5 architecture (Lecun et al., 1998) before adjusting hyperparameters to tune based on results and literature. The parameters we were adjusting were the number of layers (convolutional, fully connected etc), kernel size, number of filters, pooling type and size, dropout rate, optimisation algorithm, activation function, stride and padding. All of these were adjusted in parallel depending on results and whether the model was underfitting, overfitting etc. These adjustments were made manually during the training process.

Starting with the number of layers, kernel size and number of filters, we tested values the following values for kernel sizes, 3, 6, 9 and 16. The following number of filters were tested 8, 16, 32 and 64. The final number of layers was 5, consisting of 2 convolutional layers and 3 fully connected layers. The main problem we faced developing this model was overfitting. Along with reducing the complexity of our architecture, we also utilised dropout layers and regularisation. We placed a dropout layer at the end of each convolutional block with a dropout rate of 0.5. dropout ranges of 0.25 to 0.6 were tested. Additional dropout layers between the fully connected layers were also tested however this started to cause underfitting. The optimisation algorithms tested were Adam and RMSProp, Adam performed the best and is commonly used throughout the literature. The ReLu activation was selected because it is computationally efficient, it has been shown to work well in many applications, such as image classification and object detection and it can help prevent vanishing gradients. After 300 epochs, the model was no longer improving and the iteration with the lowest validation loss was used. A stride of 1 was used to maintain a high spatial resolution of the feature map and capture more local features.

The final model architecture was implemented using the Keras functional API (Ketkar, 2017) and can be seen in Figure 6.2. The model had an input shape of with a depth of 9 (1 ECG channel, 1 GSR channel and 7 Eye channels) and a height of 300 (300 timesteps equal to 5 seconds). The first layer is a 1D Convolutional layer with a kernel size of 9 and a filter size of 32. The max-pooling layer was then used with a pooling size of 2. The proceeding dropout layer had a dropout rate of 0.5. The next convolutional block had the same values as the previous apart from the input shape with a height of 149. Three fully connected layers with 100, 50 and 25 neurons proceed the convolutional layers. The output layers a SoftMax layer with 4 units. The test and validation accuracy and loss throughout the training epochs can be found in Figure 6.3.

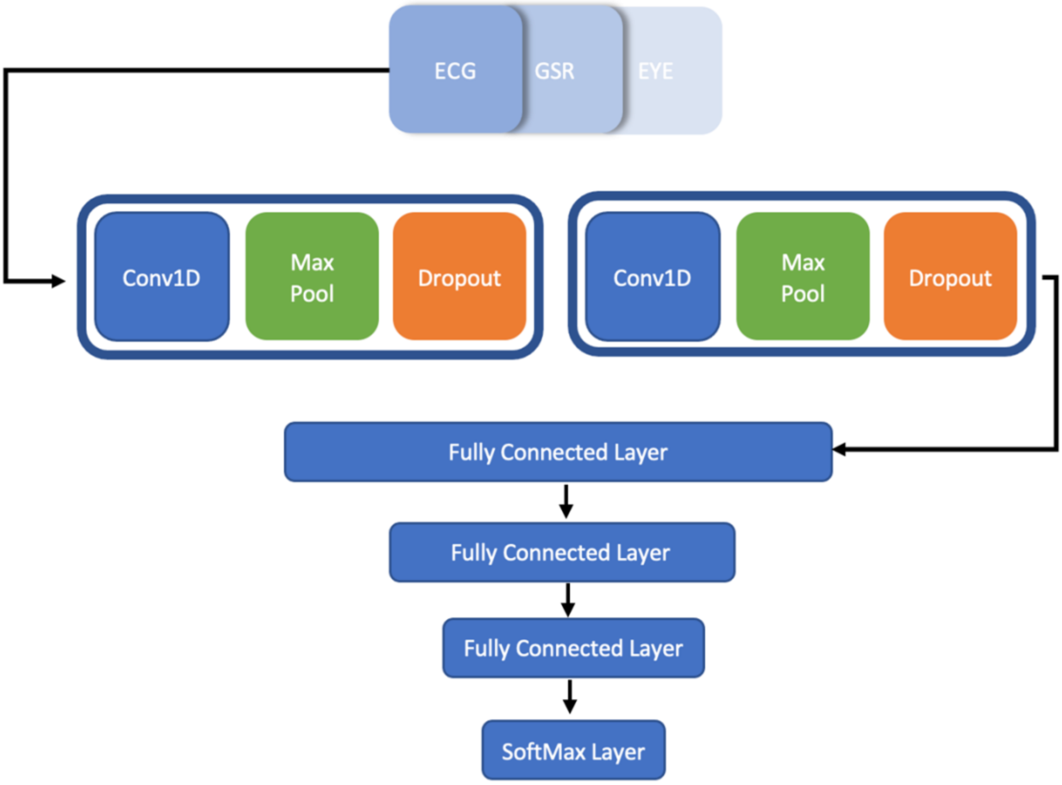


Figure 6.2: CNN Architecture

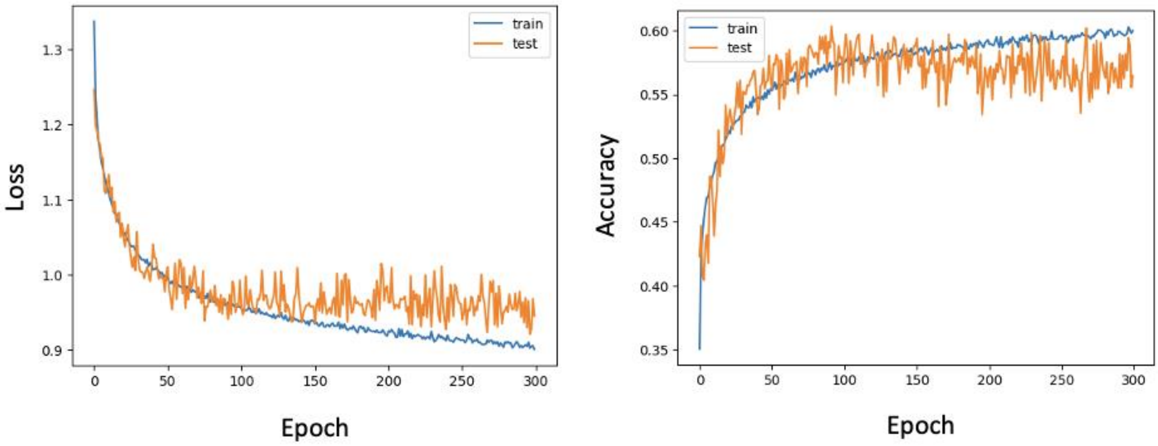


Figure 6.3: CNN training and test accuracy and loss

6.2.2.2 LSTM Models

A recurrent neural network (RNN) is a type of neural network designed to process sequential data. RNNs have a loop structure that allows them to retain information from previous time steps and use it to make predictions about future events in the sequence. RNNs are commonly used for time series data. LSTM is a type of RNN that is able to process long sequences of data and retain information for long periods of time. LSTMs are particularly useful for tasks that involve predicting a sequence of events, such as language translation and language modelling.

LSTMs have proven to be very effective for a range of tasks involving sequential data, including language translation, language modelling, and speech recognition (Sherstinsky, 2020). They are able to effectively capture long-term dependencies in the data and maintain a long-term memory of the input sequence. These reasons make LSTM a logical choice for our time series data. LSTMs have already been used to detect emotion using physiological signals.

In section 5.2.2.1 we mentioned that using ECG signals to detect arrhythmias was a task where the use of machine learning and deep learning in particular is currently being researched a lot. LSTMs along with CNNs are also producing state-of-the-art results, one example of this uses a bi-directional LSTM to classify atrial fibrillation and normal sinus rhythm and the gait of patients with Parkinson’s disease and healthy controls highlighting the versatility of LSTMs (Pham, 2021). An LSTM has been utilised to analyse ECG and GSR signals in tandem with temperature to detect emotion (Zitouni et al., 2023). The results gained using an LSTM are superior to baseline results reported in K-EmoCom (Park et al., 2020) that utilise “shallow methods”. This particular LSTM had 2 bi-directional LSTM layers with 20–50 hidden units (different number of hidden units used for various tests). A dropout of 0.2 was used after each LSTM layer. These layers fed into a fully connected layer with 100 units.

There is also research that demonstrates the effectiveness of LSTMs when using eye-tracking data to detect emotion. Eye movement features such as saccade fixation duration etc was used with EEG data to detect emotion in participants under sleep deprivation (Tao & Lu, 2020). There was a 4-layer LSTM for each modality, they all combined and fed into a fully connected layer. Before the predictive layer. A bimodal LSTM has also been used to achieve state-of-the-art results on the DEAP (Koelstra et al., 2012) and SEED (Zheng et al., 2019) datasets (Tang et al. 2017). Each node of the bimodal LSTM contains 2 LSTM layers with dropout, L2 regularisation and 8 to 64 hidden units. These hyperparameters presented above will be used to inform the development of our LSTM model.

6.2.2.2.1 LSTM Development

Our hyperparameter tunning was guided by the literature and results on the validation data. We tuned the number of LSTM layers, the number of hidden units, whether the LSTM layers were bidirectional or not, the number of fully connected layers and their hidden units, the dropout rate, regularisation and the activation function. These adjustments were made manually during the training process.

Increasing the number of LSTM layers predictably caused overfitting even when the number of hidden units was lowered. The number of Hidden units tested was 3, 6, 10, 16 and 32. This could be considered low, however, our model started to overfit with 16 hidden units even when dropout was applied. A dropout rate of 0.5 was applied to the LSTM layer to counteract overfitting. A bidirectional LSTM layer was used over a unilateral LSTM layer, it achieved better results in our testing and was common throughout the literature. Our final model had 1 bidirectional LSTM layer with 10 hidden units and a dropout rate of 0.5. This fed into a fully connected layer with 25 hidden units and a SoftMax layer was the output layer. After 200 epochs, the model was no longer improving and the iteration with the lowest validation loss was used. The model was implemented using the Keras functional API (Ketkar, 2017). The training and validation accuracy loss can be found in Figure 6.4.

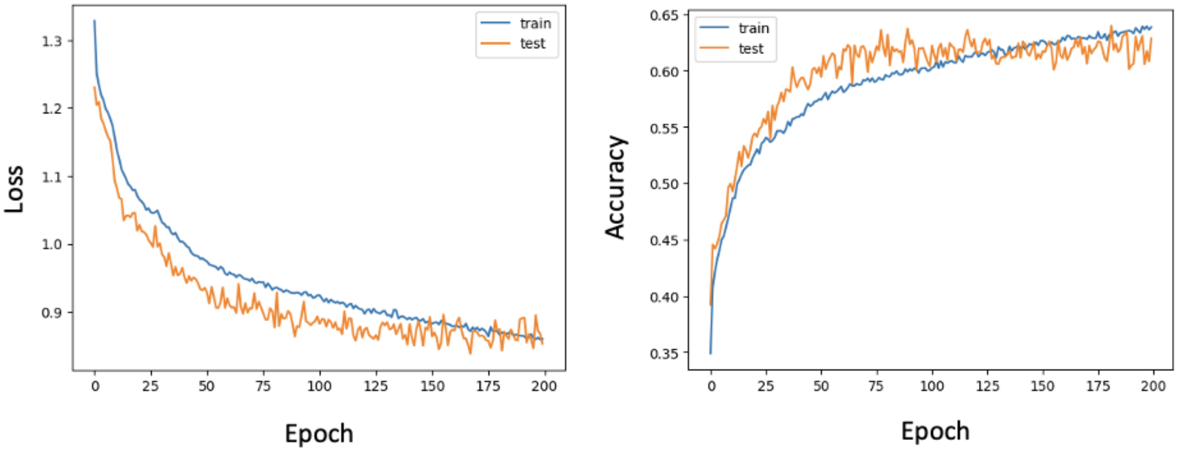


Figure 6.4: LSTM training and validation accuracy and loss

6.2.2.3 LSTM-CNN Hybrid Model

A CNN-LSTM network is a type of network architecture that combines CNNs and LSTMs. The theory behind combining a CNN and LSTM is that the unique strengths of both can be combined and lead to improved performance. The CNN component excels at extracting spatial features from the input data, while the LSTM component excels at modelling temporal features.

CNN-LSTM models have been used to analyse ECG, GSR and Eye Tracking data for many different purposes. Like individual CNNs and LSTMs, hybrid models have been used to predict cardiac arrhythmias (Rai et al., 2020; Petmezas et al., 2021). Hybrid models have also been used to analyse ECG signals for many other tasks such as disease recognition (Liao et al., 2021) and detection of QRS complexes in noisy data (Yuen et al., 2019). Social authentication is a use case where GSR signals have been analysed with success (Ekiz et al., 2021). CNN-LSTMs have been used with eye-tracking data for tasks such as early eye gaze intent prediction (Koochaki & Najafizadeh, 2019) and driver stress prediction (Mou et al., 2021).

In addition to the tasks already mentioned, hybrid models have been used extensively in emotion detection and have provided state-of-the-art results. One example of these studies utilises a CNN-LSTM to outperform baseline results on the DREAMER (Katsigiannis & Ramzan, 2018) and AMIGOS (Santamaria-Granados et al., 2019) datasets (Dar et al., 2020). This model utilised a CNN-LSTM to analyse their ECG and GSR data and a CNN for the EEG data. The CNN-LSTM consisted of two 1D convolutional layers with dropout and ReLu activation. These fed into a 128 unt LSTM and 2 fully connected layers with 256 and 128 units. Another study demonstrates that a CNN-LSTM can outperform a standard CNN or MLP for emotion detection (Kanjo et al., 2019). This research detects emotion while users are walking around a city using heart rate, GSR, body temperature, motion (accelerometer and gyro), environmental such as noise levels, UV, air pressure and GPS location data. This study uses two 2d convolutional layers with 32 and 64 filters and 2x2 max pool kernels followed by an LSTM layer. The use of CNN-LSTM to detect emotion using eye-tracking data is limited, however, there are studies that utilise eye-tracking video data (Setianto et al., 2022). This research demonstrates that a CNN-LSTM can outperform CNN and LSTM.

6.2.2.3.1 CNN-LSTM Hybrid Model Development

The development of our CNN-LSTM model commenced by building upon the initial architectures of the individual CNN and LSTM models. Subsequently, we focused on streamlining and enhancing the model's performance. The CNN component remained unchanged from the original single CNN model (please see section 5.2.2.1.1 for more details). On the other hand, the LSTM section underwent modifications. It now incorporates a single bi-directional LSTM layer consisting of 8 hidden units, accompanied by a dropout rate of 0.25. In comparison, the single LSTM model employed 10 hidden units. Additionally, we introduced a fully connected layer comprising 16 units, which serves as input to a SoftMax layer. After 200 epochs, the model was no longer improving and the iteration with the lowest validation loss was used. The model was implemented using the Keras functional API (Ketkar, 2017). The training and validation accuracy loss can be found in Figure 6.5.

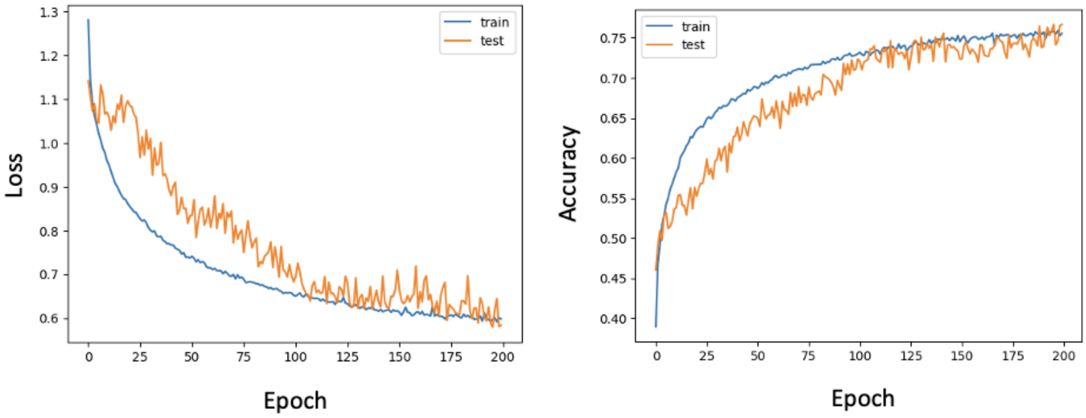


Figure 6.5: CNN-LSTM training and validation accuracy and loss

6.2.3 XAI Methodology

XAI allows AI systems to have increased transparency, trust, acceptance, fairness, compliance and less bias. It is a requirement for the responsible and ethical deployment of AI technologies. Producing a system that can be trusted, have ways to identify bias and adhere to legal requirements that demand transparency and accountability is critical when developing systems for the medical and healthcare domain. For these reasons, we provide a methodology to explain the predictions of our model.

Our goal is to offer visual explanations of predictions that have practical applications. By utilising these visualisations, we aim to deepen our comprehension of the key features that the model relies on for its predictions. This approach will enable us to enhance our understanding of the model itself, leading to improved insights and knowledge.

To accomplish our objectives, we will utilise the SHAP framework, which provides a method for interpreting the output of ML models. Proposed by Lundberg and Lee in 2017, SHAP is based on game theory (Štrumbelj and Kononenko, 2014). The framework employs the Shapley value, which measures the average marginal contribution of a feature (in our case, feature refers to varying lengths of time series data) to the model's predictions across all possible combinations of features. This calculation considers the number of ways each combination can be formed. It is worth noting that the importance of certain features, such as the occurrence of an R peak, is not dependent on the specific time step, and it is the importance of the R peak that would provide value to the system's user, not the timestep. Therefore, we cannot provide the global importance of these features. We have utilised SHAP (equation 2), which calculates the values on the predictions directly, providing computationally efficient and meaningful feature importance independent of the time step.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right].$$

$F = set\ of\ all\ features$
$\mathrm{S} \subseteq \mathrm{N} \setminus \{\mathrm{i}\}$ $represents\ each\ possible\ subset\ of\ features\ that\ does\ not\ include\ feature\ i$
$|\ \mathrm{S}\ |$ $is\ the\ number\ of\ players\ in\ subset\ S$
$\mathrm{v(S)}$ $is\ the\ value\ of\ coalition\ S, as\ defined\ by\ the$ characteristic function v
$\mathrm{v(S} \cup \{\mathrm{i}\}) - \mathrm{v(S)}$ $represents\ the\ marginal\ contribution\ of\ feature\ ii\ to\ the\ coalition\ S$

(2)

In order to implement SHAP in our specific use case, we needed to create our own implementation rather than rely on existing libraries. While there are libraries available for integrating SHAP with various ML models, none of them proved suitable for our requirements, prompting us to develop a custom implementation tailored to our needs.

For local explanations, we adopted a strategy of splitting the time series data into segments, which we consider individual features. This segmentation allows us to analyse the importance of each feature in contributing to the model's predictions. By calculating SHAP values for all these segmented features, we gain insights into the specific impact of each feature on individual predictions.

In our analysis, we focus on the absolute SHAP values, which take into importance whether they contribute positively or negatively to the prediction. Our primary interest lies in highlighting the most significant factors that influence the prediction. By examining the absolute SHAP values, we can discern the most crucial aspects of the data that drive the model's decisions.

To facilitate understanding and interpretation, we visualise the signal data while highlighting the importance of each feature. This visual representation allows for a quick and intuitive overview of the segments within the signal structure that hold the greatest importance for a specific prediction. Through these visualisations, we gain valuable insights into the specific areas of the raw time series data that significantly impact the model's predictions.

To the best of our knowledge, our implementation represents the first application of SHAP to visualise the importance of multimodality raw time series data in relation to predicting emotion. This approach enables us to provide novel insights and contribute to the advancement of interpretability and explain ability in the context of time series analysis.

To this end, we will examine both global (modality importance) and local explanations in order to identify common themes and provide reasoning behind the predictions made by the model. Firstly, we will analyse global explanations of individual modalities to determine the importance of each modality across all predictions. This analysis will offer valuable insights into the overall significance of both the model and each modality, allowing us to understand what to specifically focus on in the subsequent local explanations.

Next, we will delve into local explanations for each modality, aiming to gain a comprehensive understanding of which structural elements within a signal are crucial for predictions. This investigation will help us ascertain whether these important elements are specific to certain classes or apply universally across all predictions. By examining these local explanations, we can identify the specific components of the data that contribute significantly to the model's decisions.

A random subsampling methodology to estimate the prevalence of common themes within the data is used. This approach involves selecting a random subset of samples (n=100 per person, 25 per class) equally balanced between the 4 predicted classes, from the overall dataset. By analysing this subset, we aim to derive a representative estimate of the proportion of samples that exhibit each identified theme. To quantify the reliability of these estimates, we will calculate confidence intervals, providing a statistical range within which the true proportion for the entire dataset is likely to lie. Specifically, we will employ a 90% confidence level. Specifically, the Wilson score interval (Wilson, 1927) is used.

The overall goal of these analyses is to gain meaningful insights. By exploring global and local explanations, we can determine whether the model relies on features that have already been established as important in the medical literature. This would instil confidence in the system, as it aligns with existing knowledge. Furthermore, if the model utilises signal segments that are not widely recognised as significant, it could lead to new discoveries.

6.3 Results

6.3.1 Deep Learning Results

In this section, we present the results obtained from our model development process, including the CNN, LSTM, and hybrid CNN-LSTM models. We proceed by comparing the performance of these models and selecting the best-performing one. Furthermore, we explore the impact of different window sizes and compare the majority vote results against the baseline results achieved in Chapter 5. Additionally, we provide visualisations of the time series classification for each trial of every test participant in the testing data.

6.3.1.1 CNN Results

In this report, we present the outcomes of our CNN model. The CNN demonstrated an average accuracy, precision score, recall score, and f1 score of 0.566, 0.524, 0.575, and 0.524, respectively, across the three participants (see table 6.1). Notably, the model exhibited consistent performance for all participants, with accuracies falling within a narrow range of 3.4%. However, it is worth mentioning that Class 3 (HA/LV) and Class 1 (LA/LV) proved to be the most challenging class to predict accurately for all three participants. Surprisingly, Label 1 (LA/LV) consistently faced misclassification as Class 3 (HA/LV), indicating that instances of low arousal and low valence were consistently misclassified as instances of high arousal. This observation contradicts the baseline results, which indicated higher accuracies when predicting arousal compared to predicting valence.

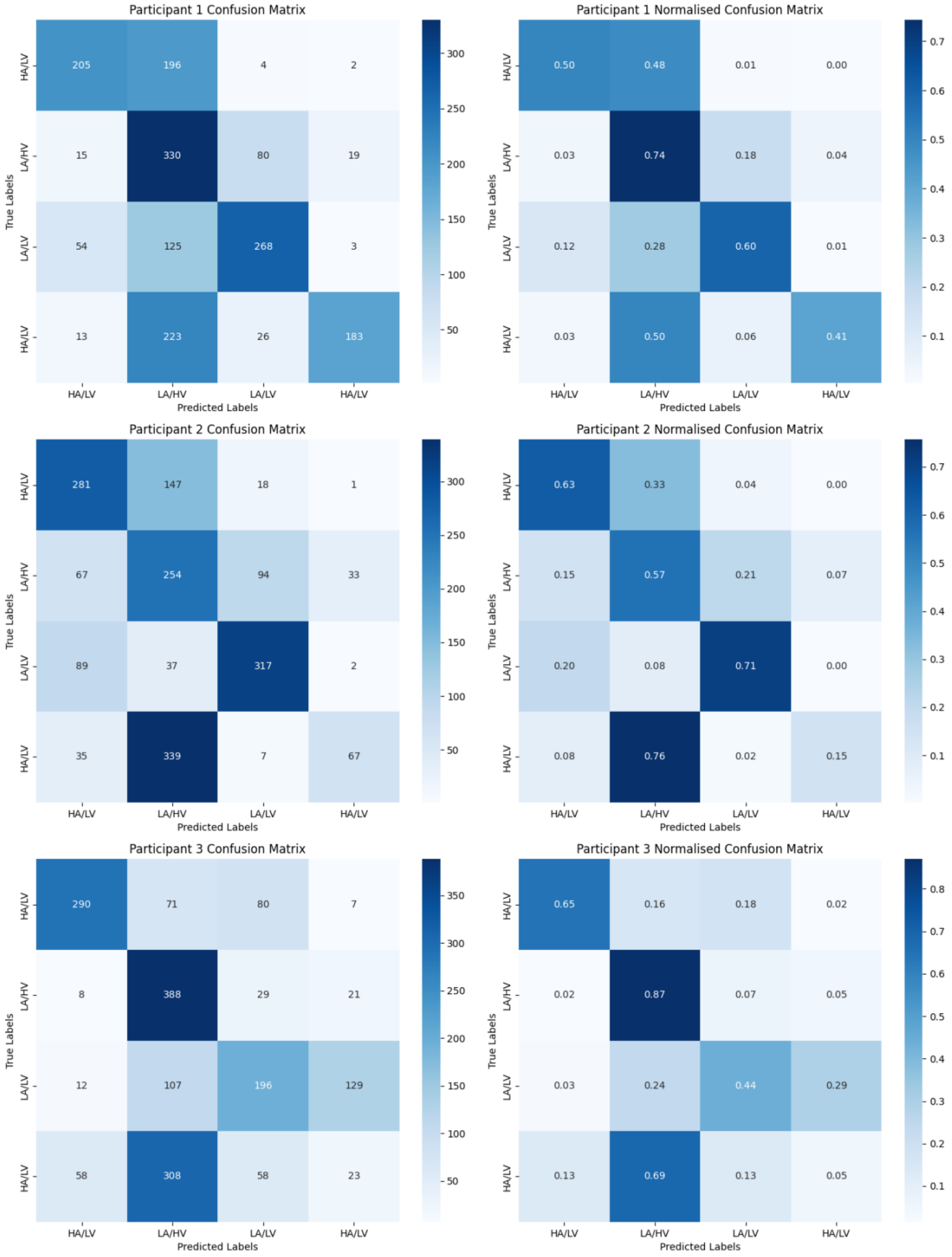


Figure 6.6: CNN confusion matrix for Participant 1, Participant 2 and Participant 3

Table 6.1: CNN Results

	Accuracy	Precision	Recall	F1
Validation	0.555	0.545	0.560	0.552
Participant 1	0.586	0.552	0.652	0.574
Participant 2	0.552	0.516	0.560	0.506
Participant 3	0.560	0.504	0.513	0.494
Combined Participants	0.566	0.524	0.575	0.524

6.3.1.2 LSTM Results

The LSTM model demonstrated an average accuracy, precision score, recall score, and f1 score of 0.546, 0.521, 0.610, and 0.512, respectively, across the three participants (see table 6.2). In contrast to the CNN, the LSTM exhibited a larger range of accuracies, with a difference of 10.3% between the highest and lowest values and a standard deviation of 0.045. Similar to the CNN, the LSTM also encountered challenges in accurately predicting Class 3 (HA/LV) for all three participants. Additionally, Class 1 (LA/HV) was consistently misclassified as Class 3 (HA/LV), indicating a recurring issue in distinguishing instances of low arousal and low valence from instances of high arousal. Notably, for Participants 1 and 2, the LSTM predominantly predicted Class 0 (HA/LV) and Class 1 (LA/HV), respectively, suggesting a bias towards these classes in the model's predictions.

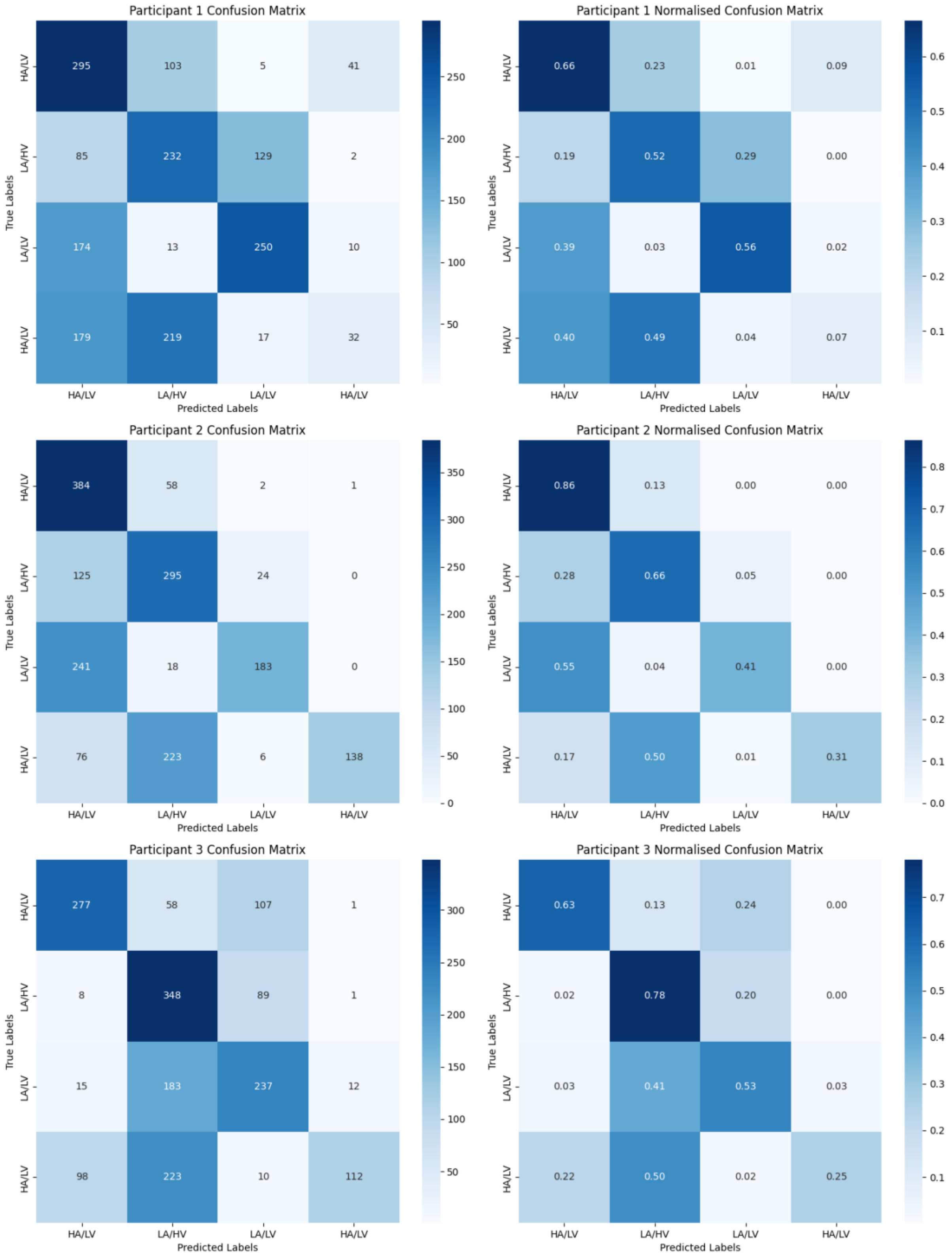


Figure 6.7: LSTM confusion matrix for participant 1, participant 2 and participant 3

Table 6.2: LSTM Results

	Accuracy	Precision	Recall	F1
Validation	0.612	0.590	0.600	0.595
Participant 1	0.482	0.453	0.458	0.424
Participant 2	0.571	0.564	0.717	0.554
Participant 3	0.585	0.547	0.656	0.557
Combined Participants	0.546	0.521	0.610	0.512

6.3.1.3 CNN-LSTM Results

The CNN-LSTM model achieved an average accuracy, precision, recall, and f1 score of 0.631, 0.630, 0.632, and 0.621, respectively (see table 6.3). The accuracies showed a range of 14.1% and a standard deviation of 0.061. Interestingly, Class 3 (HA/LV) was frequently misclassified as Class 1 (LA/HV) in participants 2 and 3, but not in participant 1. The performance in this aspect was significantly better than using either the single CNN or LSTM models alone. However, participant 3's results were notably worse compared to participants 1 and 2. Additionally, we present the Majority Vote results, where the entire video was classified as a whole (n=12) instead of segmenting it into 5-second intervals. The majority vote accuracy was 74.3%, which will be compared to the baseline results presented in chapter 4, we will also provide the precision, recall, and f1 scores for the majority vote classification.

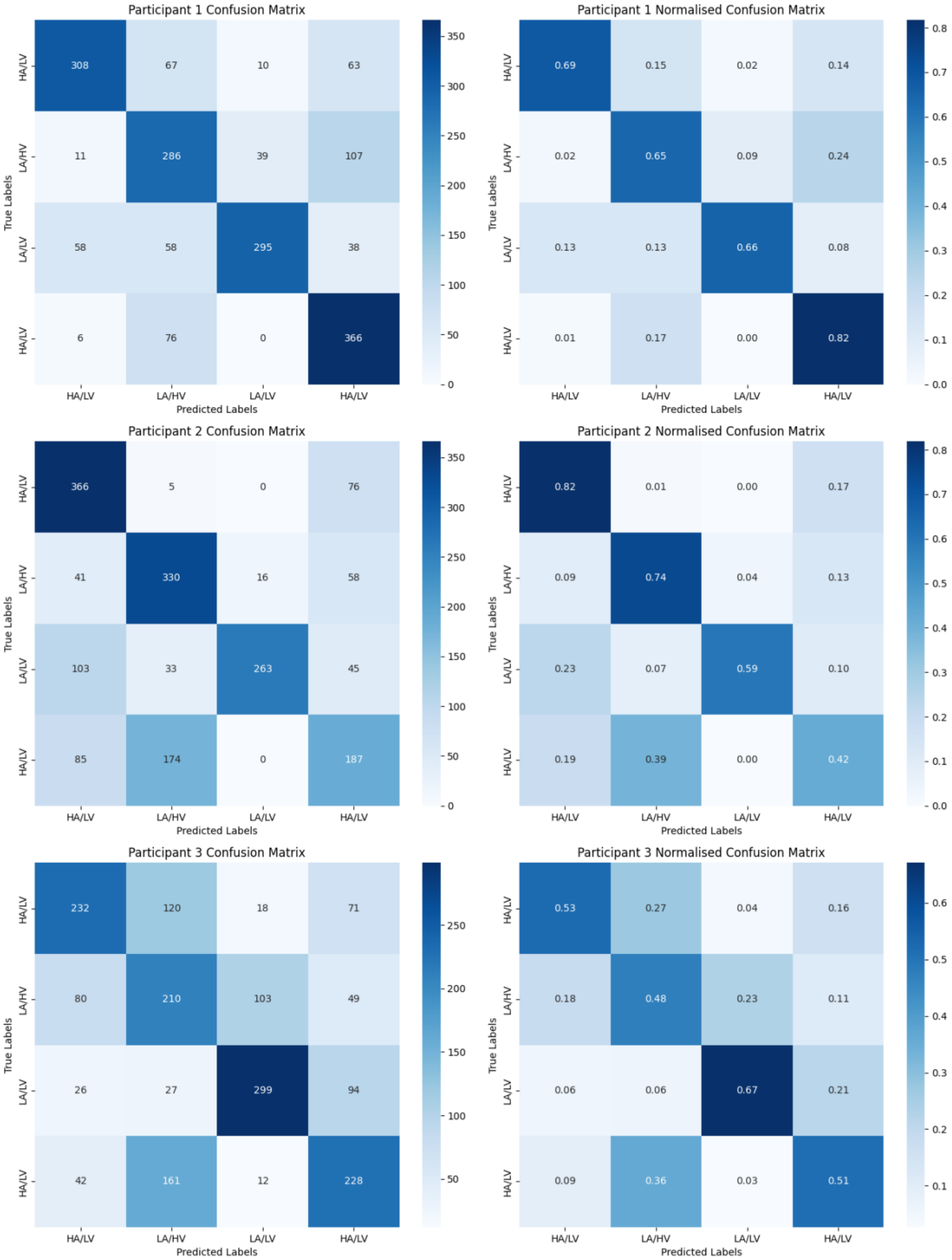


Figure 6.8: CNN-LSTM (5-second window) confusion matrix for Participant 1, Participant 2 and Participant 3

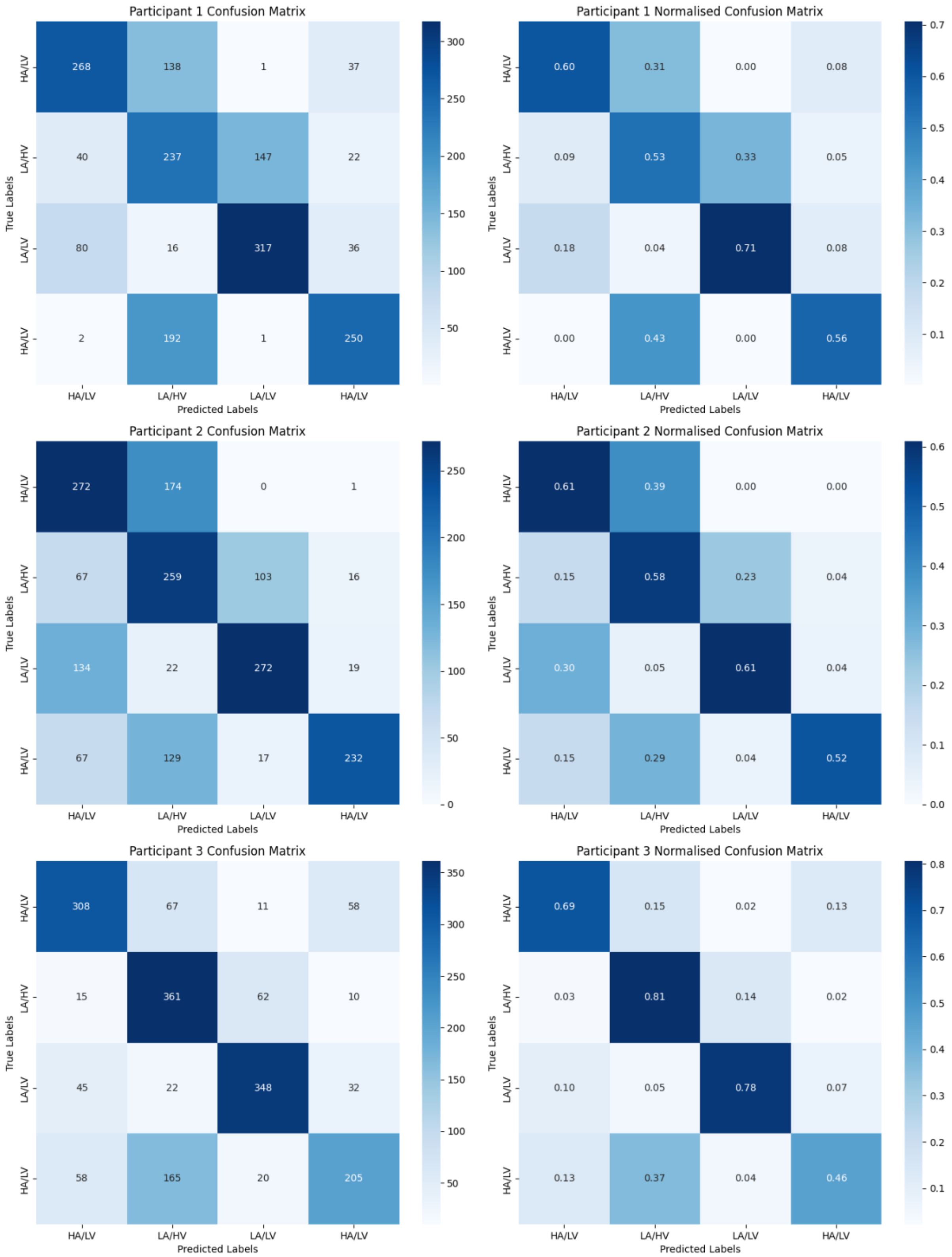
Table 6.3: CNN-LSTM results

	Accuracy	Precision	Recall	F1	Majority Vote
Validation	0.751	0.700	0.742	0.720	N/A
Participant 1	0.687	0.702	0.700	0.690	0.833
Participant 2	0.660	0.646	0.659	0.636	0.818
Participant 3	0.546	0.543	0.537	0.538	0.583
Combined Participants	0.631	0.630	0.632	0.621	0.743

6.3.1.3 Results Comparison

6.3.1.3.1 Window Size Results

In this analysis, we compare the performance of different window sizes for our classification task. The 2.5-second window achieved accuracy, precision, recall, and f1 scores of 0.634, 0.622, 0.642, and 0.627, respectively (see table 6.4). On the other hand, the 10-second window resulted in lower scores of 0.424, 0.396, 0.419, and 0.385 for accuracy, precision, recall, and f1, respectively (see Table 6.5). It is evident from the results (see Table 6.6) that both the 2.5-second and 5-second windows outperform the 10-second window. While there is a debate on whether the 2.5-second or 5-second window is better, we have decided to proceed with the 5-second window. The 2.5-second window shows slightly better real-time accuracy, recall, and f1 scores, but the 5-second window significantly improves majority vote predictions.



Participant 1 Confusion Matrix

Participant 1 Normalised Confusion Matrix

Participant 2 Confusion Matrix

Participant 2 Normalised Confusion Matrix

Participant 3 Confusion Matrix

Participant 3 Normalised Confusion Matrix

Figure 6.9: 2.5 Second window Confusion matrix for participant 1, participant 2 and participant 3

Table 6.4: 2.5 second window results

	Accuracy	Precision	Recall	F1	Majority Vote
Validation	0.723	0.680	0.699	0.689	N/A
Participant 1	0.603	0.601	0.607	0.603	0.667
Participant 2	0.586	0.581	0.620	0.589	0.636
Participant 3	0.712	0.685	0.700	0.689	0.75
Combined Participants	0.634	0.622	0.642	0.627	0.684

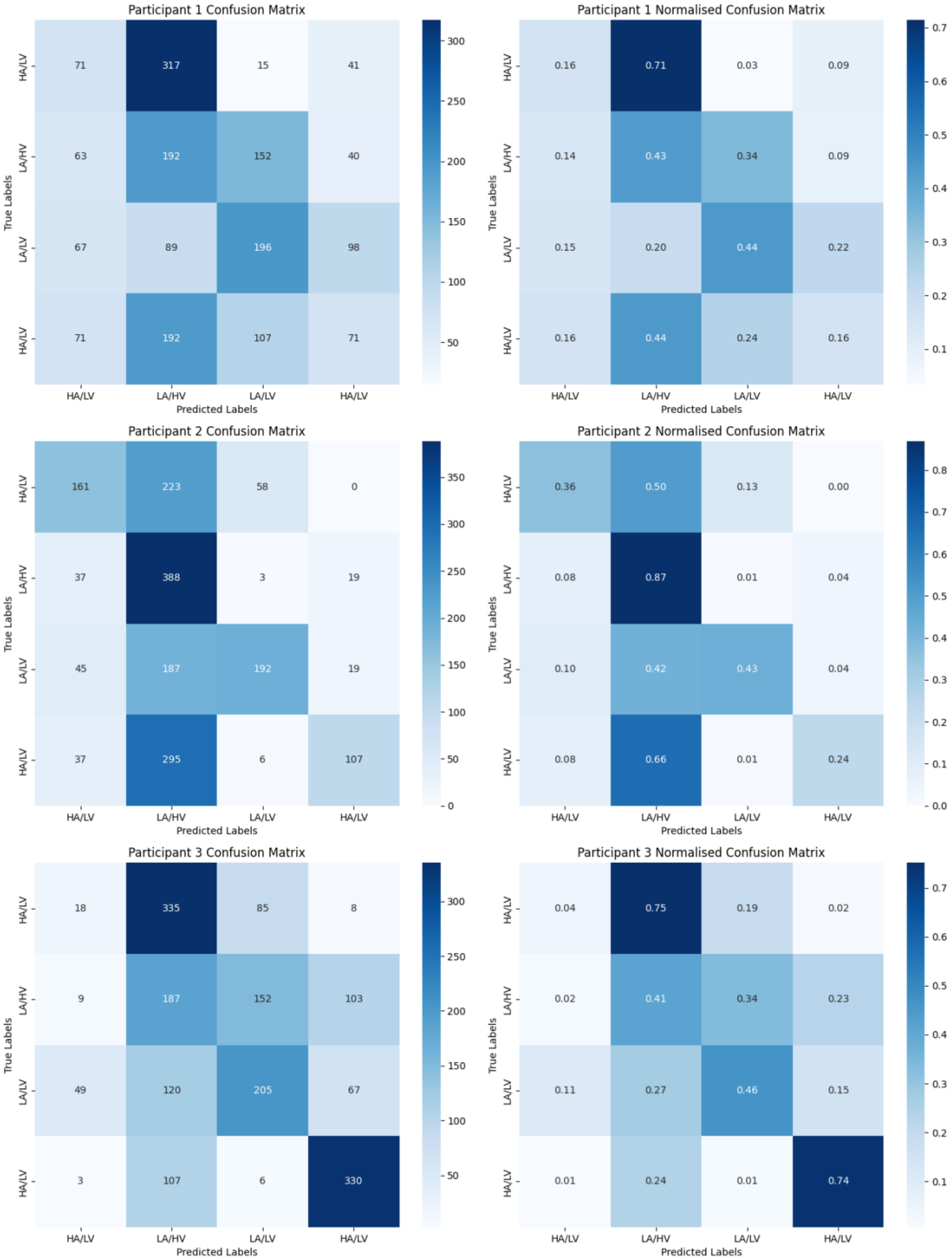


Figure 6.10: 10 Second window Confusion matrix for participant 1, participant 2 and participant 3

Table 6.5: 10 second window results

	Accuracy	Precision	Recall	F1	Majority Vote
Validation	0.585	0.621	0.557	0.587	N/A
Participant 1	0.333	0.298	0.300	0.297	0.333
Participant 2	0.531	0.475	0.590	0.482	0.545
Participant 3	0.408	0.414	0.368	0.375	0.5
Combined Participants	0.424	0.396	0.419	0.385	0.459

Table 6.6: Window Size Results Comparison

Window-size	Accuracy	Precision	Recall	F1	Majority Vote
2.5 second	**0.634**	0.622	**0.642**	**0.627**	0.684
5 second	0.631	**0.630**	0.632	0.621	**0.743**
10 second	0.424	0.396	0.419	0.385	0.459

6.3.1.3.2 Real-Time Results Comparison

In this section, we compare the results of three models: CNN, LSTM, and CNN-LSTM. As expected, the CNN-LSTM consistently outperforms all other models in every category, as highlighted in bold in Table 6.7. These findings are consistent with what has been observed in the existing literature. Given these superior performance results, we have decided to use the CNN-LSTM model to present time series classification results, visualisation, and comparisons to the baseline in future analyses.

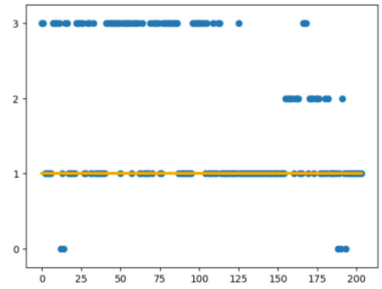
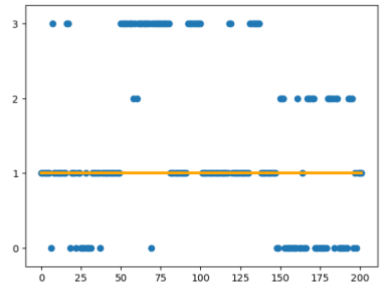
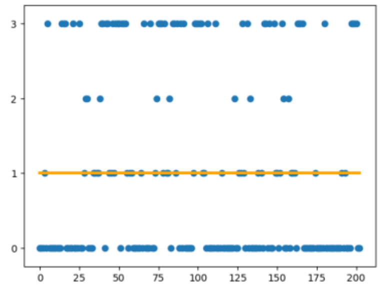
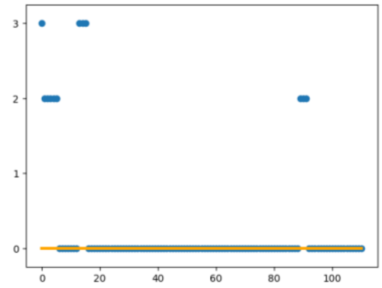
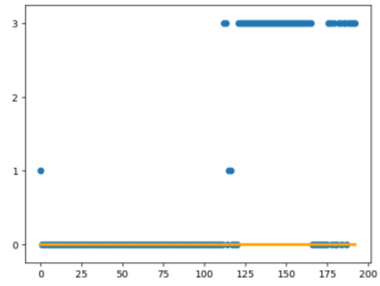
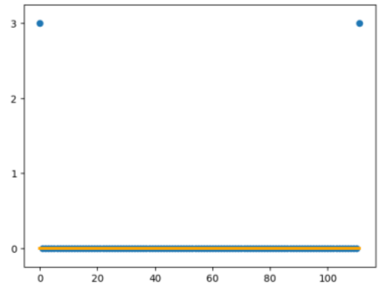
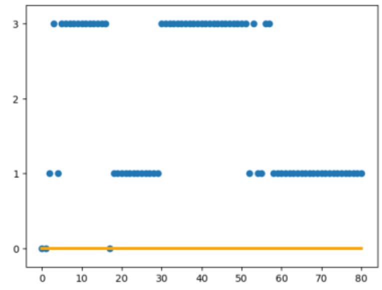
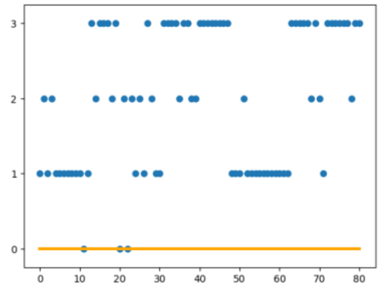
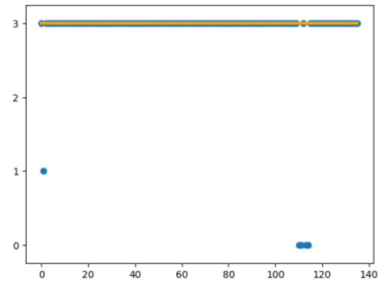
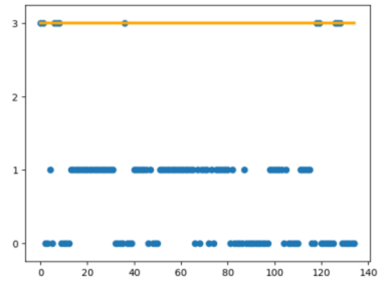
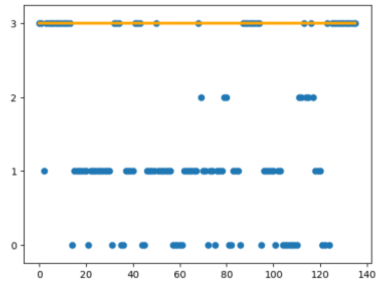
Table 6.7: Real-Time Results Comparison

Model	Accuracy	Precision	Recall	F1
CNN	0.566	0.524	0.575	0.524
LSTM	0.546	0.521	0.610	0.512
CNN-LSTM	**0.631**	**0.630**	**0.632**	**0.621**

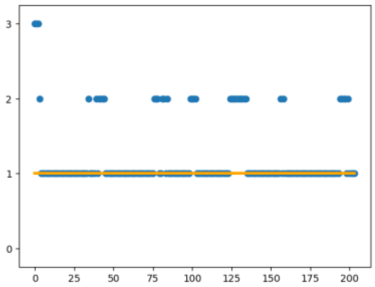
To enhance the interpretability of our findings, we provide visualisations of the time series classification for each trial conducted by every test participant in the testing dataset. These visualisations offer a comprehensive and intuitive representation of the classification outcomes, enabling us to better understand the model's performance on an individual level. By examining the patterns, trends, and potential misclassifications within the time series data, we can gain deeper insights into the model's strengths and limitations.

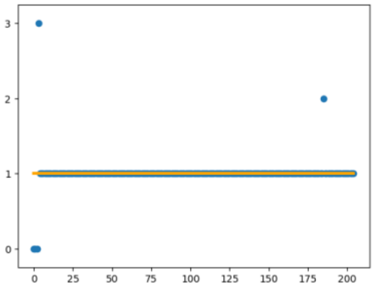
The visualisations of the data reveal several interesting themes. One initial concern was whether it would be challenging to classify the first few seconds of a video, where the initial emotions may be setting in. However, the time series classification analysis presented in Table 6.8 indicates that there are no clear indications of difficulty in classifying these initial moments. Additionally, consistent themes emerged across all participants, with videos that were classified with high accuracy demonstrating consistency across participants. Conversely, when there is a consistent misclassification of a video, this misclassification tends to be consistent across participants as well. For example, in the case of the last two videos, the majority of predictions indicate class 3, but all misclassifications occur in predicting class 1. These findings highlight the presence of consistent patterns and misclassifications across the analysed videos and participants.

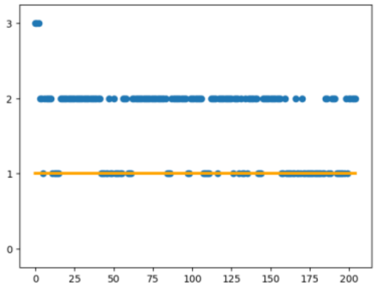
Table 6.8: CNN-LSTM Time Series Classification

Participant 1	Participant 2	Participant 3
		
Classified Correctly (1)	Classified Correctly (1)	Classified Incorrectly (0)
		
Classified Correctly (0)	Classified Correctly (0)	Classified Correctly (0)
	No data	
Classified Incorrectly (1)	N/A	Classified Incorrectly (3)
		

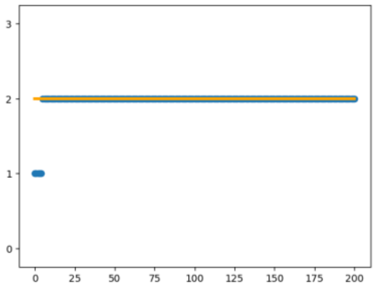
Classified Correctly (3) Classified Incorrectly (1) Classified Inorrectly (1)

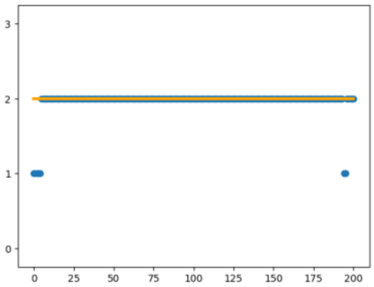


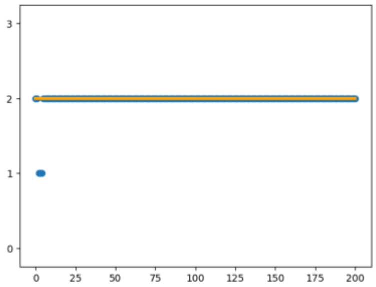




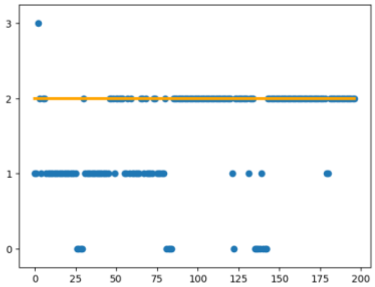
Classified Correctly (1) Classified Correctly (1) Classified Incorrectly (2)

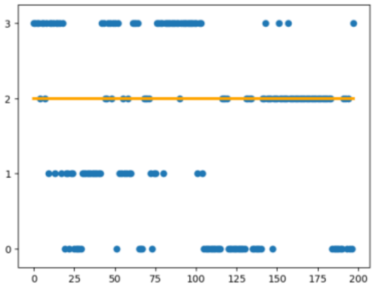


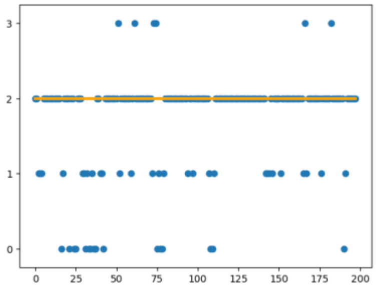




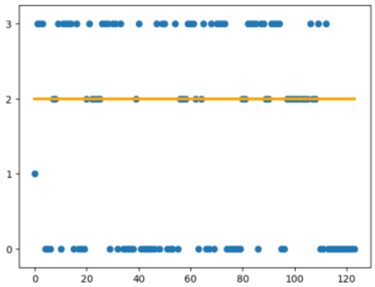
Classified Correctly (2) Classified Correctly (2) Classified Correctly (2)

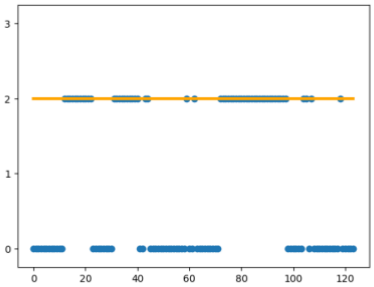


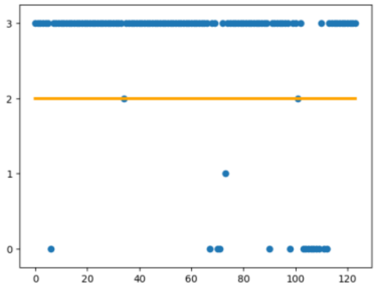




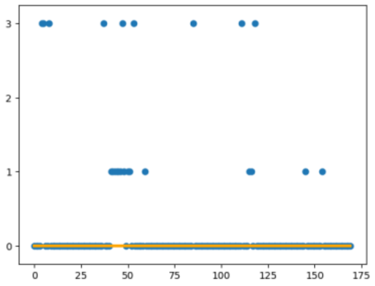
Classified Correctly (2) Classified Correctly (2) Classified Correctly (2)

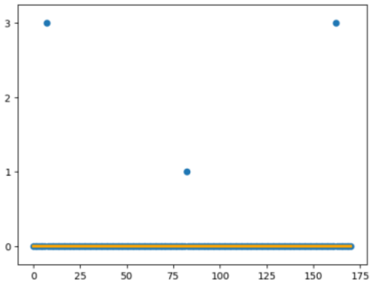


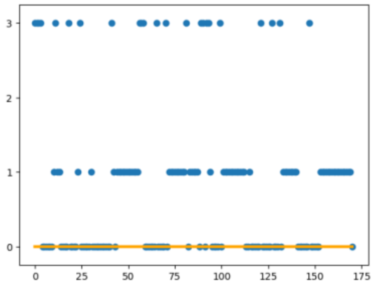




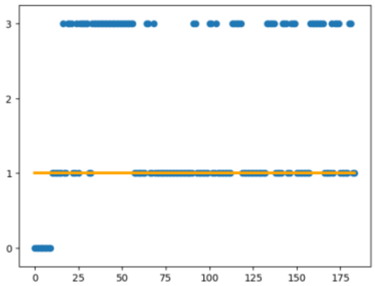
Classified Incorrectly (0) Classified Incorrectly (0) Classified Incorrectly (3)



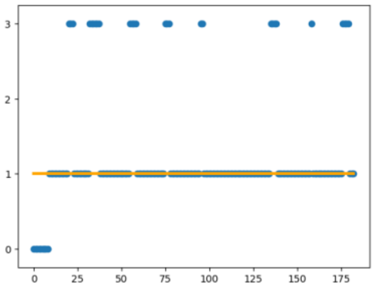




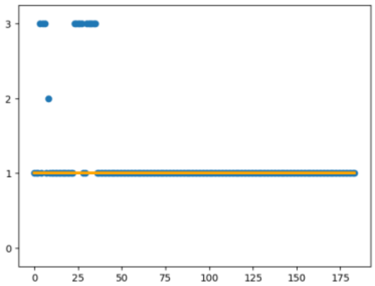
Classified Correctly (0) Classified Correctly (0) Classified Correctly (0)



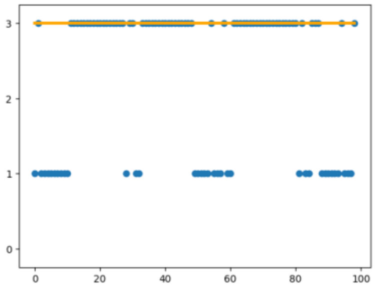
Classified Correctly (1)



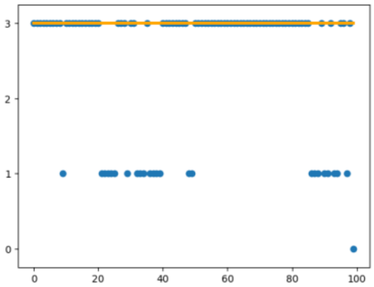
Classified Correctly (1)



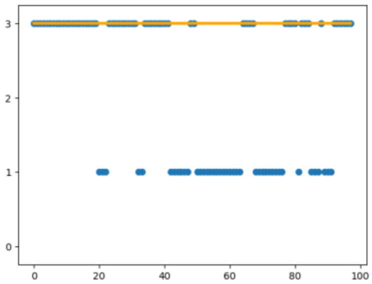
Classified Correctly (1)



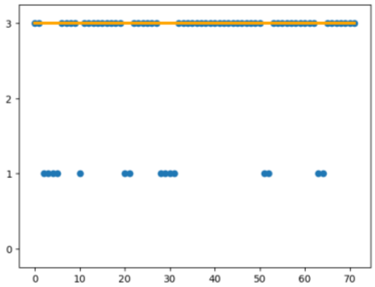
Classified Correctly (3)



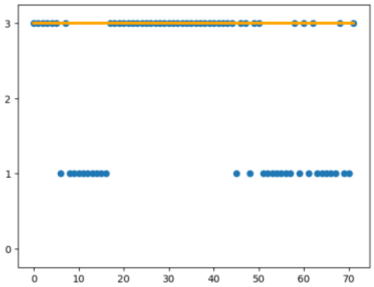
Classified Correctly (3)



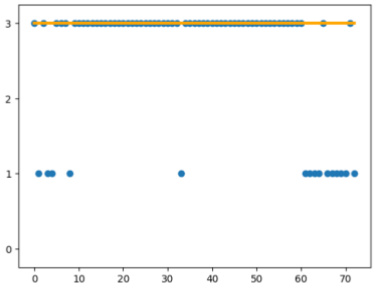
Classified Correctly (3)



Classified Correctly (3)



Classified Correctly (3)



Classified Correctly (3)

6.3.1.3.3 Majority Voting Results Comparison

To further validate our approach, we compare the majority vote results obtained from our model with the baseline results established in Chapter 4. These results are presented in table 6.9 and Figure 6.11. We used a majority vote in order to have the most direct comparison. In comparison to the baseline results, the CNN-LSTM model demonstrates superior performance across all performance metrics except precision. Examining the confusion matrix, the CNN-LSTM model showcases a more balanced outcome across all classes. Specifically, the CNN-LSTM model significantly improves the prediction results for classes 2 (LA/LV) and 3 (HA/LV). These findings highlight the CNN-LSTM's ability to outperform the baseline and provide enhanced accuracy, particularly in predicting classes 2 and 3.

Table 6.9: Majority Voting Results Comparison

	Accuracy	Precision	Recall	F1
Baseline	0.719	**0.76**	0.72	0.72
CNN-LSTM	**0.743**	0.743	**0.750**	**0.743**

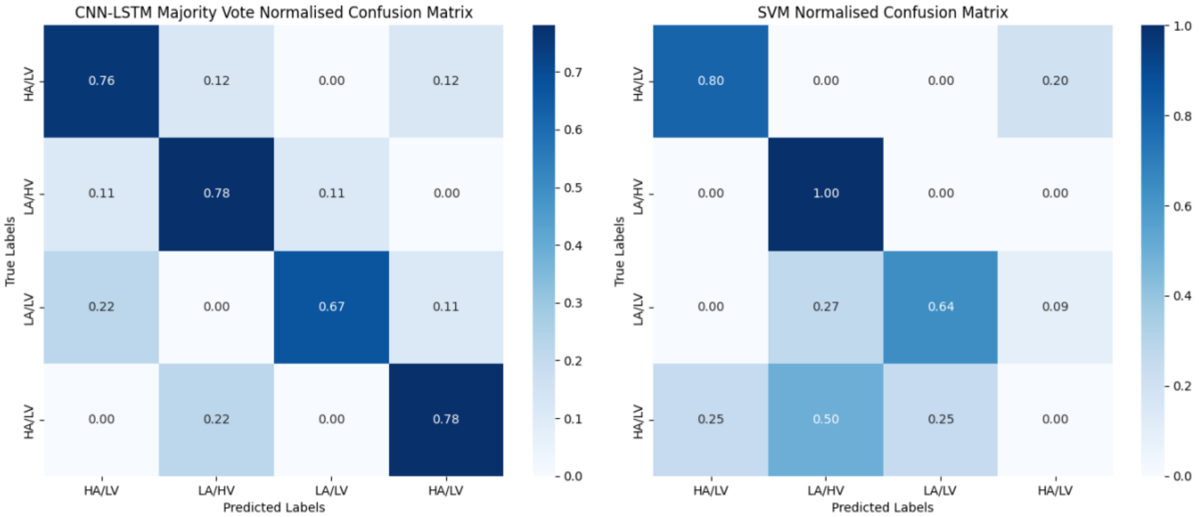


Figure 6.11: Confusion matrices for CNN-LSTM majority vote and baseline SVM results

6.3.2 XAI Results

6.3.2.1 Global Results

In this global analysis, our objective is to examine the overall modality SHAP score for each class across all participants. Our goal is to gain insights into the significance of each modality in predicting different emotional states. Additionally, we will focus on three specific eye-tracking channels, namely X, Y, and blink. By evaluating their relevance, we aim to understand their importance in predicting various emotional states. This investigation will provide valuable information regarding the contribution of each modality and eye-tracking feature to emotional state prediction.

Analysing the SHAP values, several patterns emerge. Firstly, it appears that each participant exhibits a similar pattern of importance, suggesting that the impact of modality is more dependent on the specific emotion rather than the individual participant. Secondly, when considering all the scenarios in Table 6.10, it becomes evident that ECG has the highest overall impact on predictions. ECG emerges as the most influential modality in 10 out of the 12 scenarios. Eye tracking, on the other hand, takes the second spot in terms of impact, being the most important modality in 2 out of the 12 scenarios and ranking as the second most impactful in the remaining cases. Lastly, GSR demonstrates the least impact across all scenarios in Table 6.10.

Chapter 5 presents contrasting findings in terms of the importance of different modalities for the deep learning model. Specifically, compared to the previous results, ECG demonstrates significantly higher importance in the deep learning model. This suggests a disparity between the previous findings and the current analysis. On the other hand, the importance of eye tracking remains consistent with the results presented in Chapter 5, indicating its similar level of significance. However, GSR has notably diminished impact on the predictions in the current analysis compared to the previous findings. These discrepancies emphasise the importance of reassessing and updating our understanding of modality importance in the context of the deep learning model.

There are potential explanations for the observed differences in importance across techniques. Firstly, the higher importance of ECG in the deep learning model may be attributed to the model's ability to examine specific aspects of the ECG signal that were not fully captured during the feature extraction process. To gain further insights into this, we conduct a local analysis (see section 6.3.2.2). This analysis will explore which specific parts of the ECG signal are being utilised by the model and contribute to its high importance.

On the other hand, the reduced impact of GSR could be attributed to the window size used for data analysis. It is worth noting that many of the extracted features from GSR that were identified as important are related to the number of peaks and valleys throughout the entire sample. However, due to the limited duration of the data window (5 seconds), certain characteristics of GSR signals may not be adequately captured. As a result, the impact of GSR on predictions may appear diminished compared to the previous chapter, which had longer temporal information.

A global analysis was conducted specifically on eye-tracking data, as it is the only modality with multiple channels available for analysis. The analysis focused on three channels: X, Y, and blink. As anticipated, the findings reveal that both X and Y channels have a greater impact compared to the blink channel. Moreover, the results suggest a slight marginal superiority of the Y channel over the X channel, although it is reasonable to expect a strong interrelation between them.

To gain deeper insights into the intricacies of eye movement, a local analysis was conducted. The objective of this local analysis is to explore aspects such as the speed and size of movements, which were not captured in the global analysis but are commonly considered when extracting eye-tracking features.

Table 6.10: Global analysis of modalities.

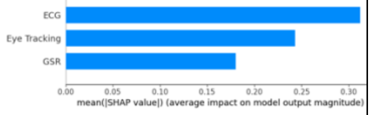
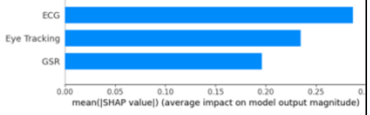
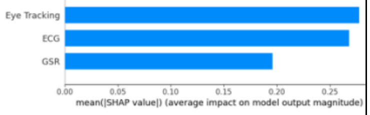
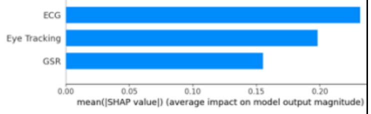
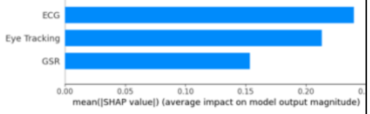
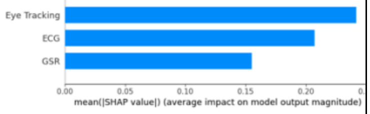
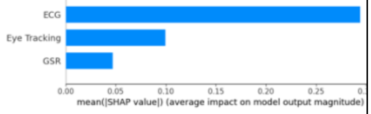
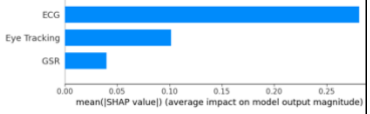
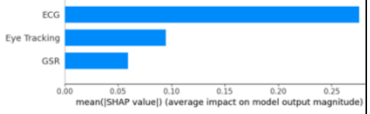
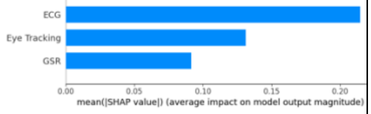
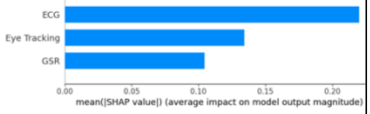
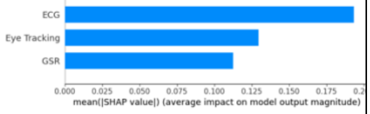
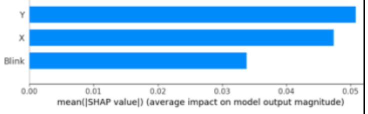
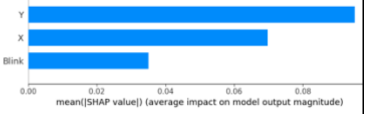
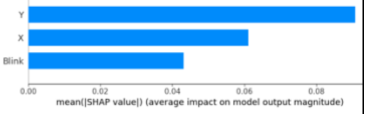
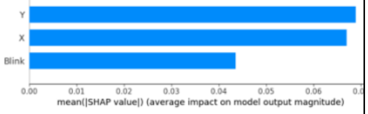
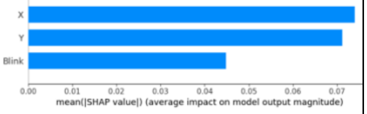
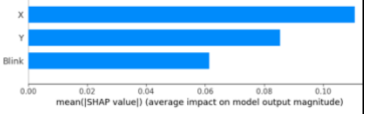
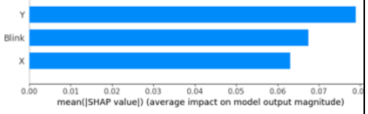
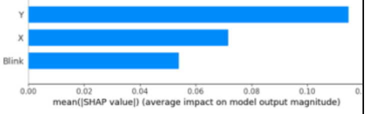
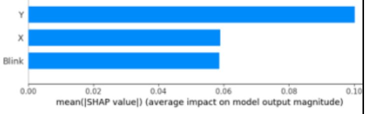
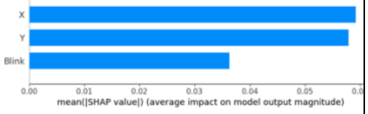
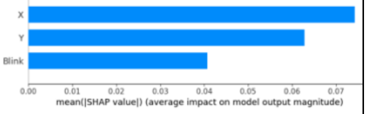
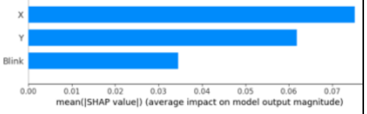
<table>
<tr><th></th><th>Participant 1</th><th>Participant 2</th><th>Participant 3</th></tr>
<tr><td>HA/HV</td><td> </td><td> </td><td> </td></tr>
<tr><td>LA/HV</td><td> </td><td> </td><td> </td></tr>
<tr><td>LA/LV</td><td> </td><td> </td><td> </td></tr>
<tr><td>HA/LV</td><td> </td><td> </td><td> </td></tr>
</table>

Table 6.11: Global analysis of Eye Tracking features

<table>
<tr><th></th><th>Participant 1</th><th>Participant 2</th><th>Participant 3</th></tr>
<tr><td>HA/HV</td><td> </td><td> </td><td> </td></tr>
<tr><td>LA/HV</td><td> </td><td> </td><td> </td></tr>
<tr><td>LA/LV</td><td> </td><td> </td><td> </td></tr>
<tr><td>HA/LV</td><td> </td><td> </td><td> </td></tr>
</table>

6.3.2.2 Local Results

During our local analysis, we thoroughly examined a variety of local examples among participants in order to identify recurring themes within the signals. This process allowed us to gain valuable insights into the common patterns that emerged. In this report, we highlight a selection of signal examples that effectively demonstrate our findings. We looked at the whole of the 5-second samples, before pinpointing the most crucial second and further dissecting it to identify the individual components of the signal being utilised by the model. This approach enabled us to extract information and draw observations from the data.

6.3.2.2.1 ECG Characteristics

In our analysis, we sought to identify overall themes that were consistent across participants and videos. Upon reviewing the data, it became evident that participants' signals generally followed similar themes throughout each video. However, there were differences observed among participants. Notably, Participant 3 consistently exhibited a lower heart rate compared to the other participants. Despite these variations, all participants shared common heart rate themes across the videos. It was observed that the HA/HV, indicating happiness and excitement, correlated with increased heart rate, while the LA/HV, representing calming and happiness, correlated with decreased heart rate. These trends align with the expected physiological responses.

In the subsequent phase of analysis, our focus shifted to identifying the most significant segment of the ECG complex. The model consistently emphasised the QRS complex, specifically the R peak, as an important indicator. This trend was observed in 72% of the cases in the random subset of data analysed, the Wilson score interval calculates that with 90% confidence, the true proportion of samples exhibiting this theme in the entire dataset lies between 67.5% and 76.1%. Additionally, the ST segment and T wave were found to be useful to the model as well, despite not being commonly highlighted in literature as direct feature extraction components. It was observed as the first or second most important component in 67% (Wilson score = 62.4%-71.3%) of the analysed samples.

Interestingly, when predicting the HA/HV and LA/HV, the model predominantly relied on the QRS complex. However, when predicting the LA/LV and HA/LV, the model extracted more information from the entire ECG complex, suggesting that a broader range of signal components was relevant for these predictions.

Table 6.12: ECG signal analysis for HA/HV

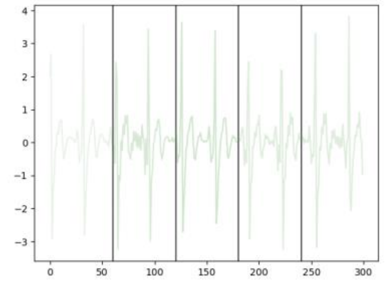
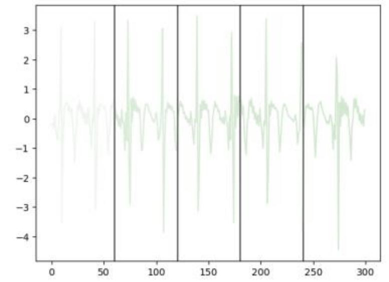
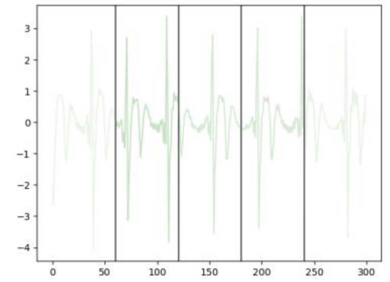
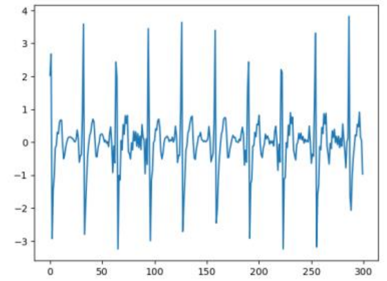
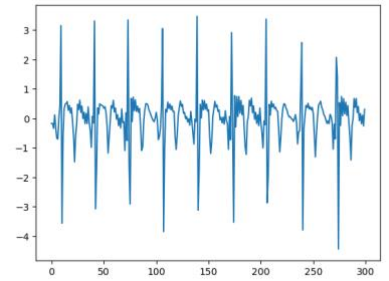
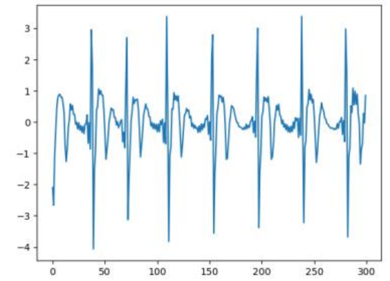
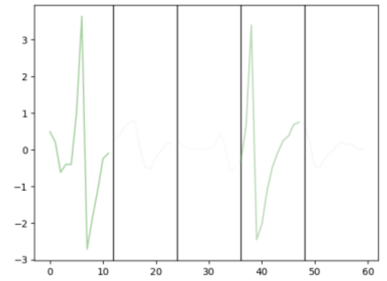
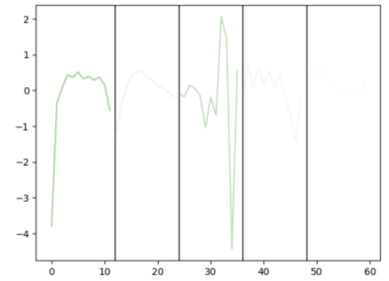
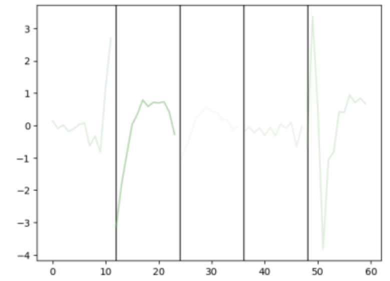
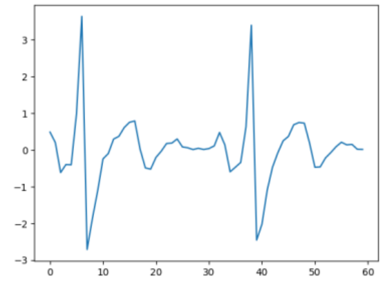
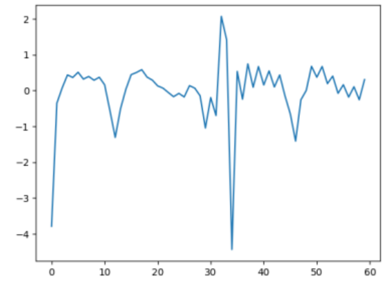
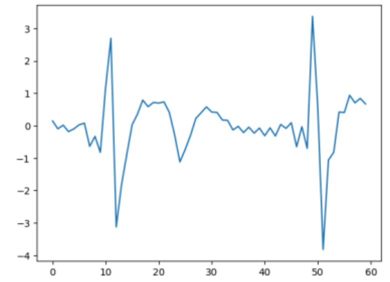
Participant 1	Participant 2	Participant 3
		
		
		
		

Table 6.13: ECG signal analysis for LA/HV

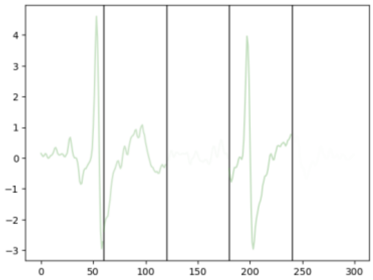
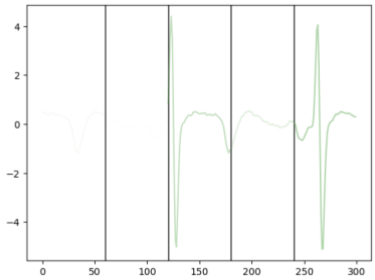
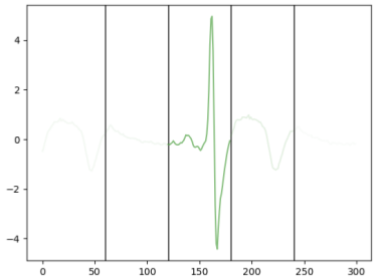
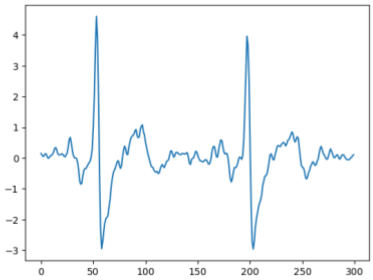
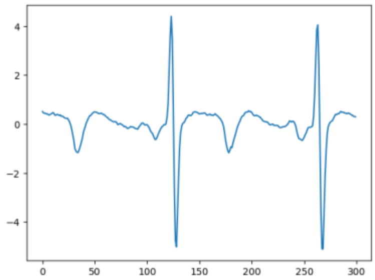
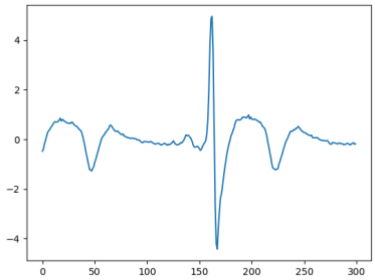
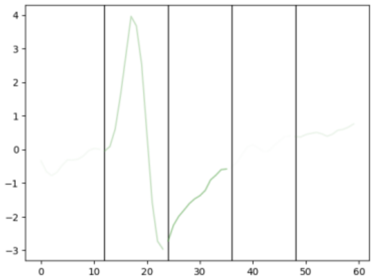
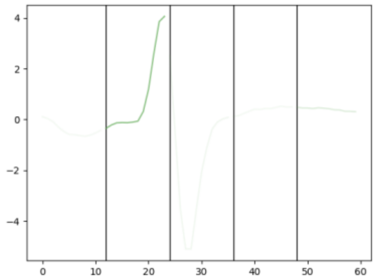
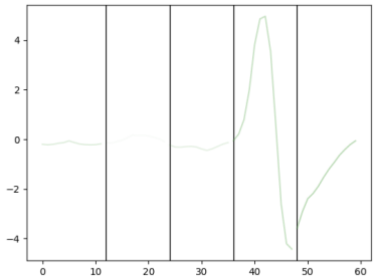
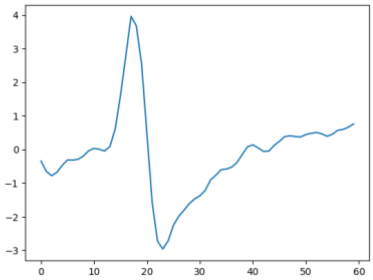
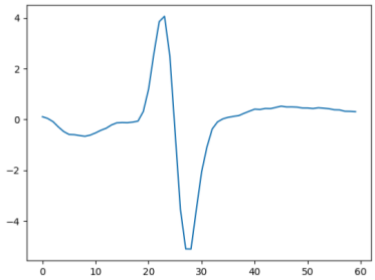
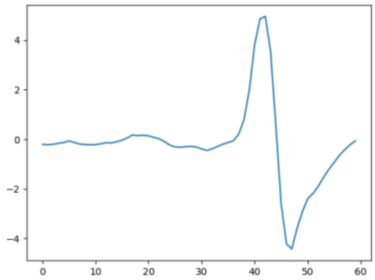
Participant 1	Participant 2	Participant 3
		
		
		
		

Table 6.14: ECG signal analysis for LA/LV

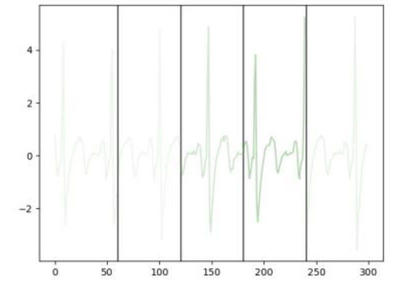
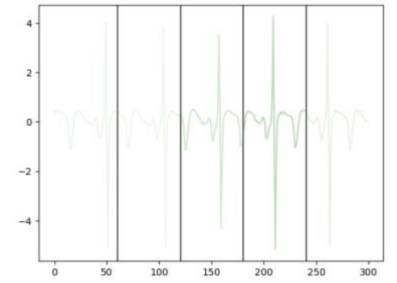
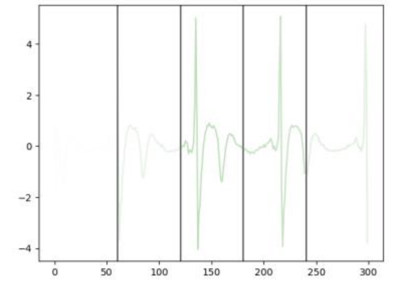
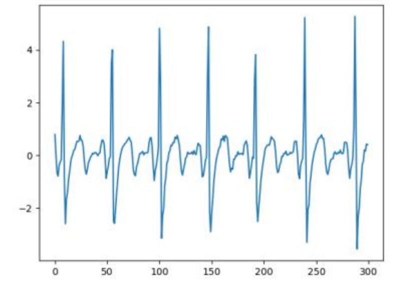
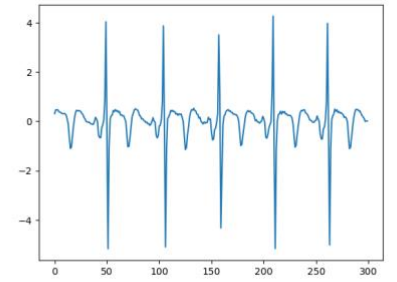
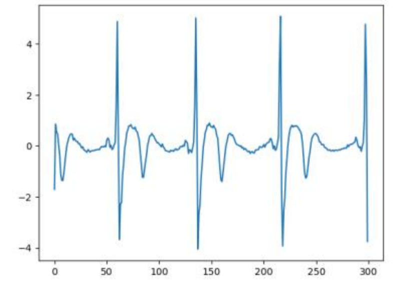
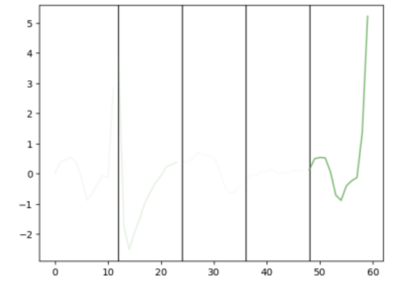
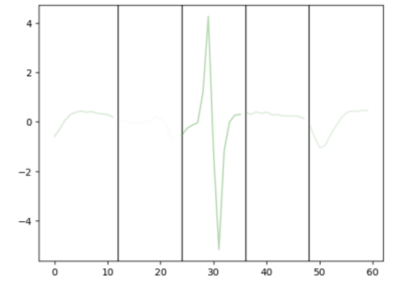
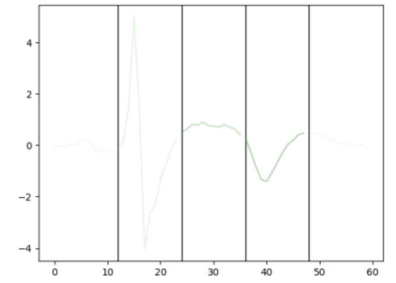
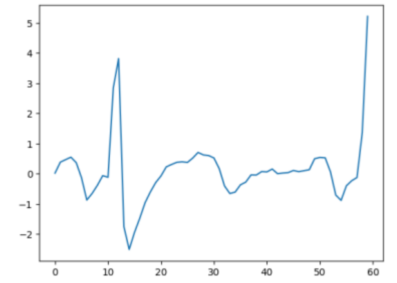
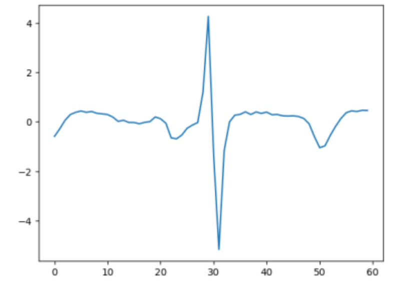
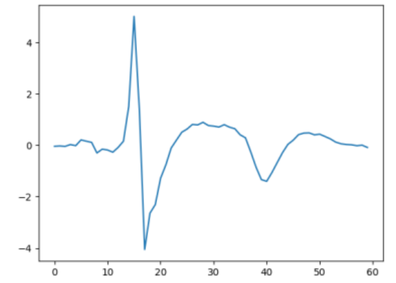
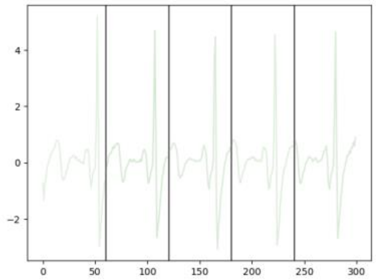
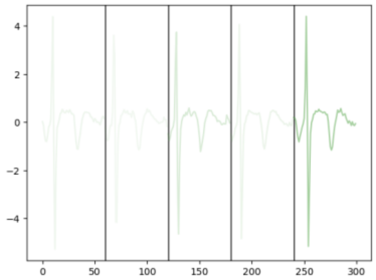
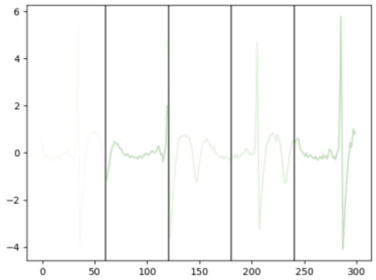
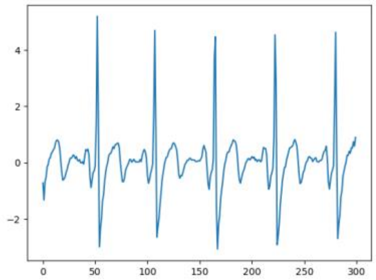
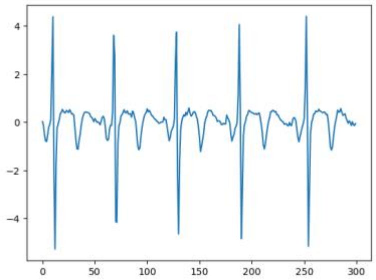
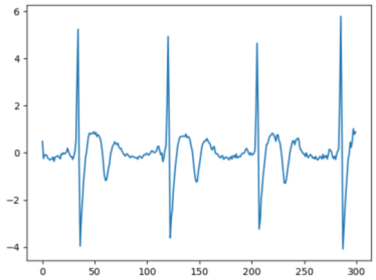
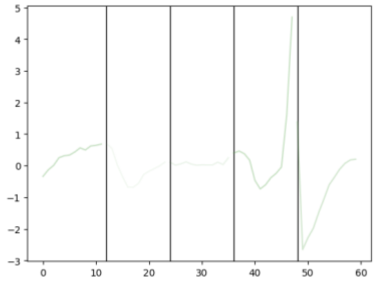
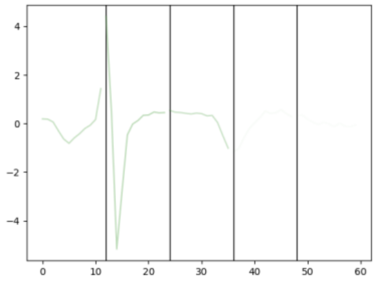
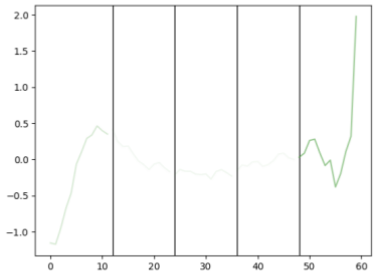
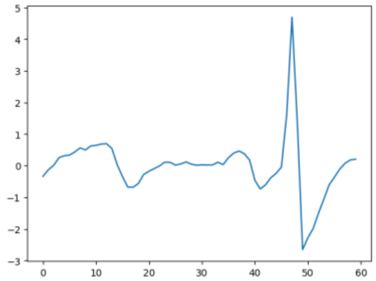
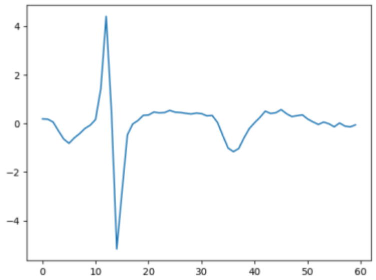
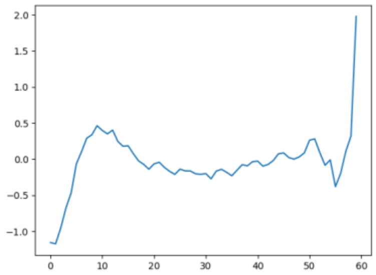
Participant 1	Participant 2	Participant 3
		
		
		
		

Table 6.15: ECG signal analysis for HA/LV

Participant 1	Participant 2	Participant 3
		
		
		
		

6.3.2.2.2 GSR Characteristics

Upon examination, we observed that examples with fairly equal importance assigned cross multiple segments tend to have higher overall importance compared to examples where one particular segment is dominant. There is a noticeable bias towards the first segments of data, the reason for this is not clear the first segment was the most important in 64% of the samples observed with an upper and lower Wilson confidence bound of 59.3% and 68.4%.

Regarding the amplitudes, HA/HV consistently demonstrates the higher amplitude across examples for all participants, while LA/LV consistently exhibits the lower amplitudes. This reinforces the importance of mean amplitude as a feature, as previously observed in chapter 4. The analysis here further confirms the relevance of this particular feature.

Although we can observe a visible trend where the most important segments are not necessarily the peaks and valleys, it remains unclear how this trend impacts the overall model. Identifying the specific themes that the model focuses on when analysing GSR signals is challenging. However, we have determined that to extract more significant information from GSR signals, a larger window size may be necessary. Increasing the window size could potentially reveal additional meaningful patterns and enhance our understanding of GSR analysis.

Table 6.16: GSR signal analysis for HA/HV

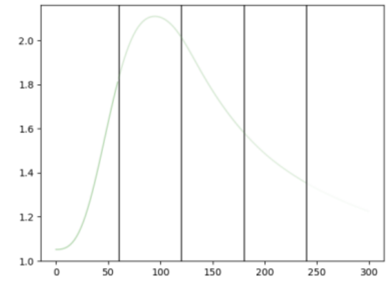
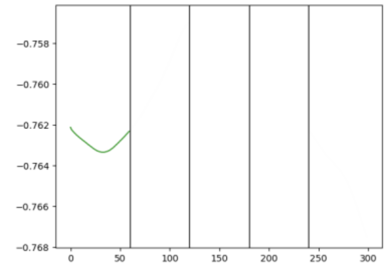
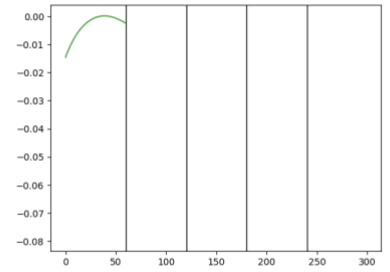
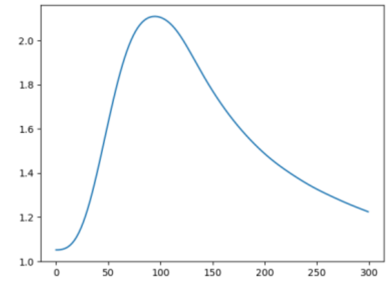
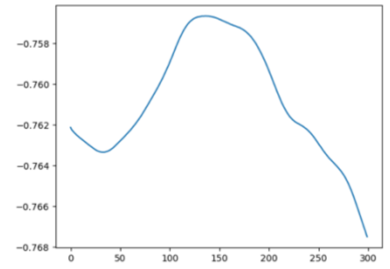
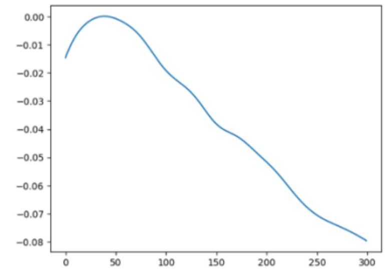
Participant 1	**Participant 2**	**Participant 3**
		
		

Table 6.17: GSR signal analysis for LA/HV

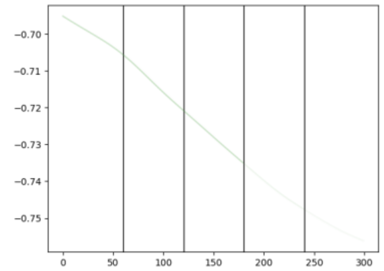
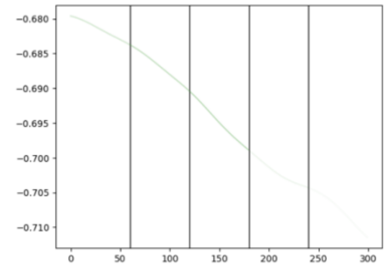
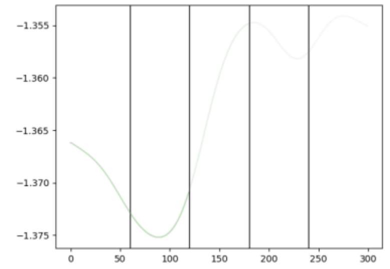
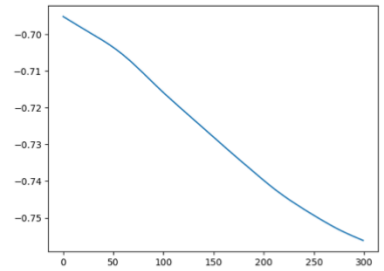
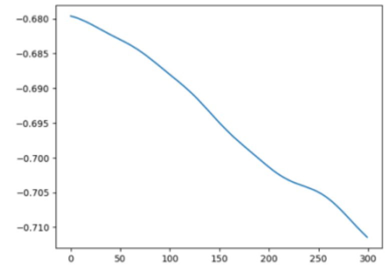
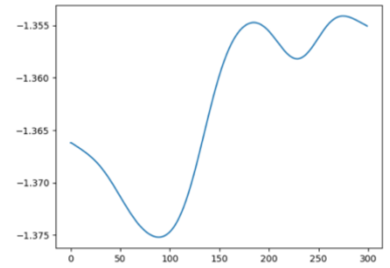
Participant 1	Participant 2	Participant 3
		
		

Table 6.18: GSR signal analysis for LA/LV

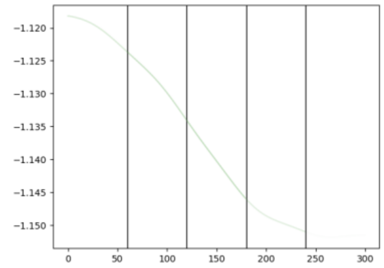
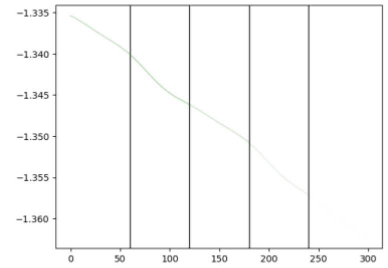
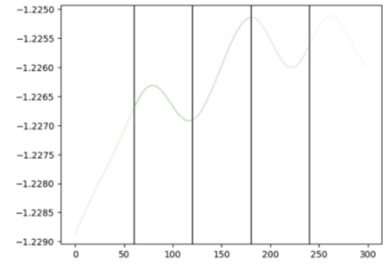
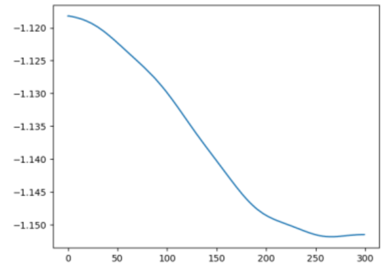
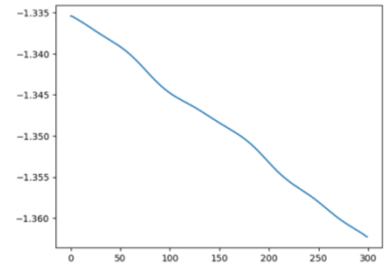
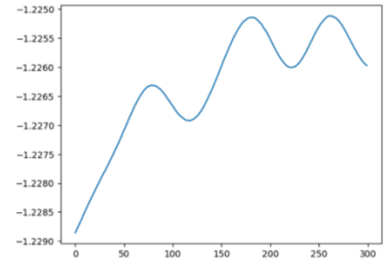
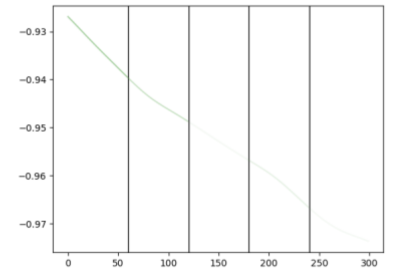
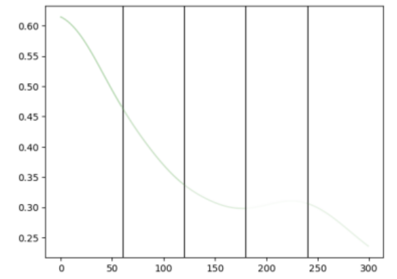
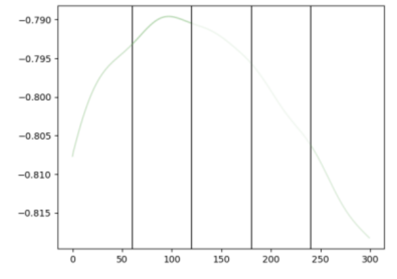
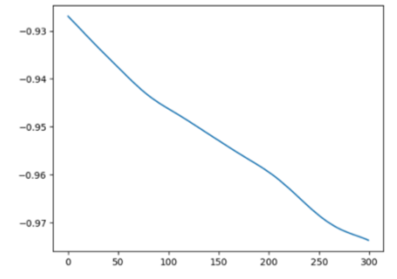
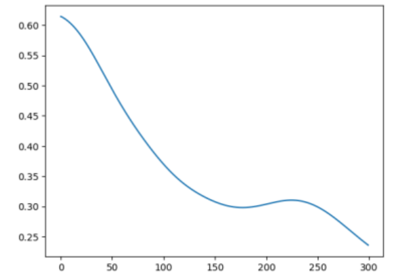
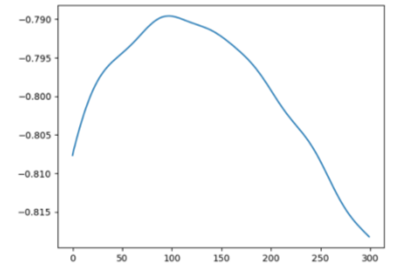
Participant 1	Participant 2	Participant 3
		
		

Table 6.19: GSR signal analysis for HA/LV

Participant 1	Participant 2	Participant 3
		
		

6.3.2.2.3 Eye Characteristics

When examining eye tracking data, we provide visual analysis consisting of a 2D map and X and Y time series data. All the examples presented in Tables 6.20, 6.21, 6.22 and 6.23 exhibit an above-average eye-tracking SHAP value.

Upon analysing the data, several key observations were made. Firstly, we found that larger and smoother eye movements, do not significantly impact the models' predictions. This contrasts with small quicker movements, this finding aligns with our initial expectations and is consistent with the extracted features and feature importance’s stated throughout literature and in Chapter 4. Micro saccades were identified as the most important eye-tracking feature in chapter 4, the sample rate we use for this analysis may not be high enough to capture these movements.

Furthermore, our global analysis revealed that the Y axis carries greater influence than the X axis, except when classifying HA/LV. Interestingly, a closer examination of the data showed that the Y data exhibited more instances of saccades compared to the X data. This disparity in saccade frequency could potentially account for the overall increased SHAP value observed in relation to the Y axis.

Overall, all segments of a given eye-tracking example contribute similar levels of importance in most cases. However, when a specific section stands out as more significant, it typically coincides with fixations accompanied by saccades. This finding suggests that moments of focused attention, indicated by fixations, play a substantial role in the overall predictive power of eye-tracking data in the case of the deep learning model. There is limited literature to compare our findings to; however, this finding is consistent with the feature importance results presented in chapter 4 and a review on emotion detection using eye tracking data (Lim et al., 2020) also highlights fixations as a commonly utilised feature.

These observations deepen our understanding of the relationship between eye movements and their impact on affective predictive models. By highlighting the role of saccades, the differential influence of the X and Y axes, the possible benefits of using data with a higher sample rate and the significance of fixations in eye-tracking data analysis, we gain valuable insights into the mechanisms underlying eye movement patterns and how deep learning models that utilise eye-tracking data can be improved.

Table 6.20: Eye-tracking analysis for HA/HV

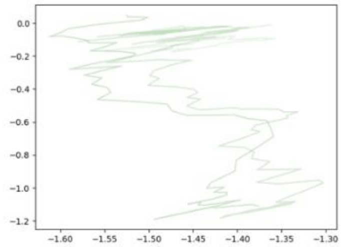
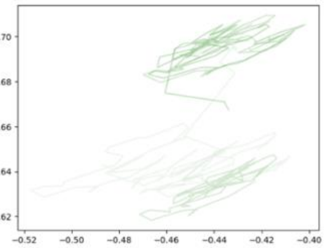
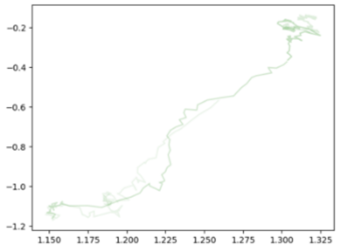
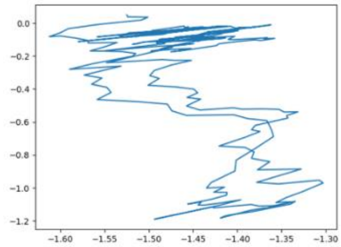
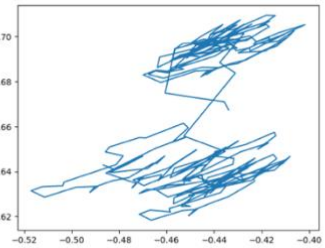
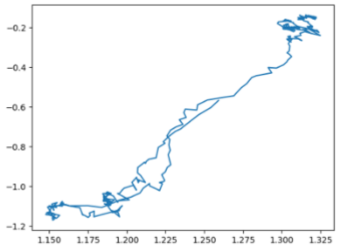
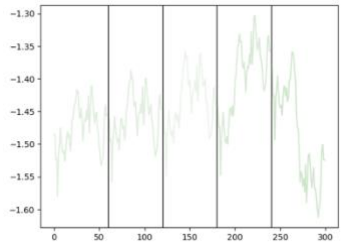
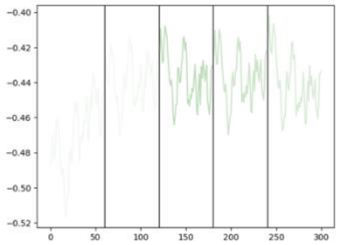
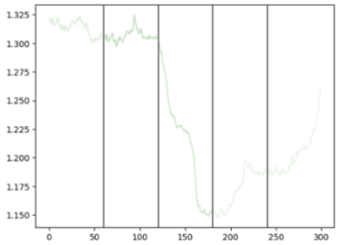
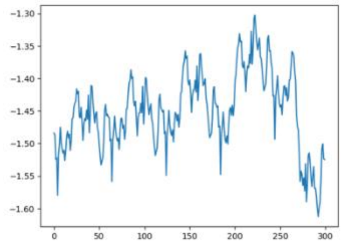
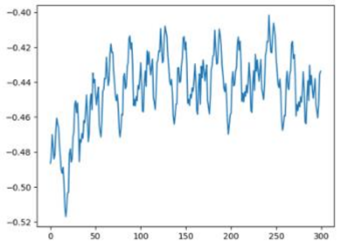
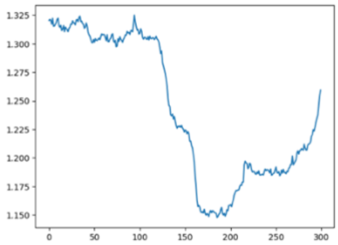
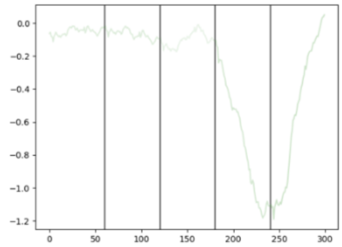
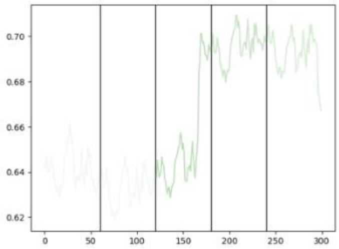
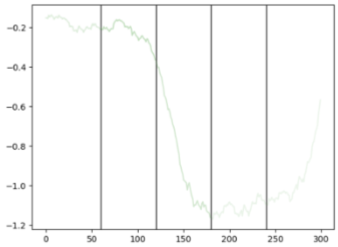
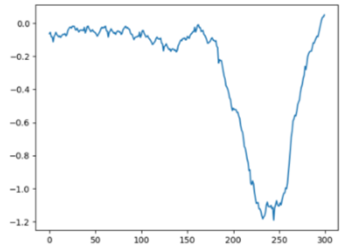
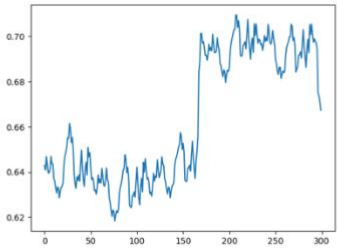
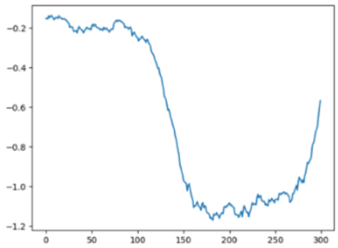
	Participant 1	Participant 2	Participant 3
2D Map			
			
X			
			
Y			
			

Table 6.21: Eye-tracking analysis for LA/HV

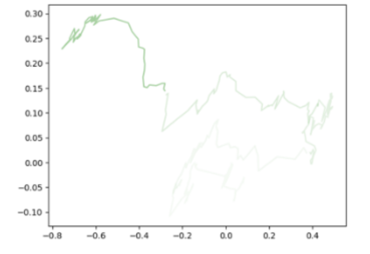
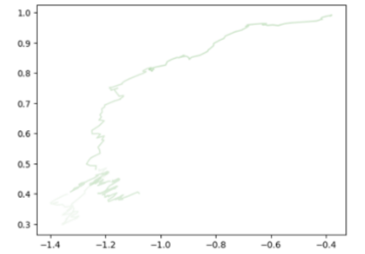
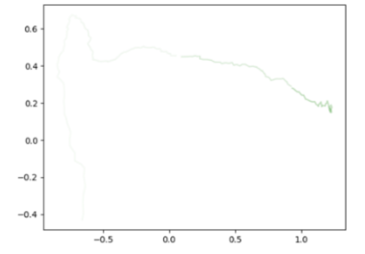
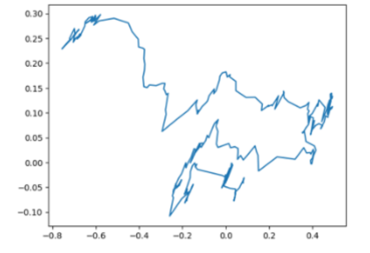
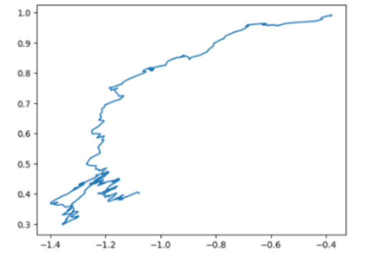
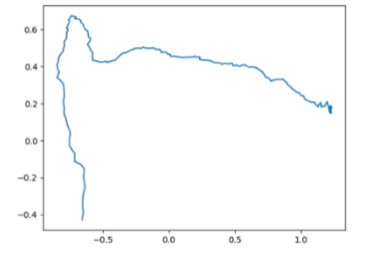
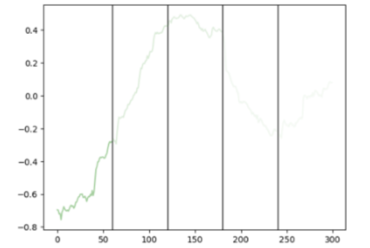
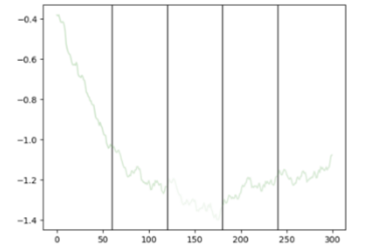
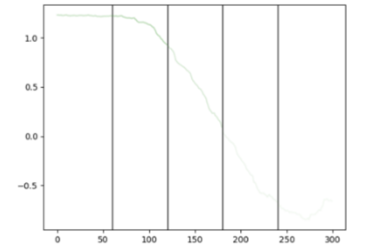
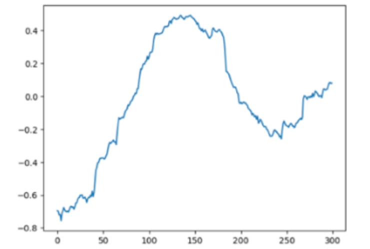
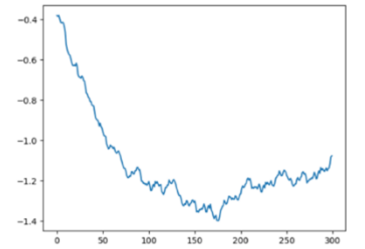
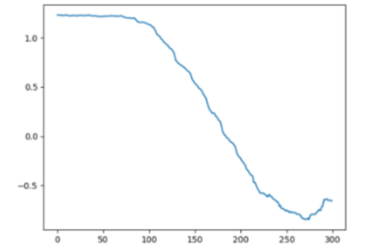
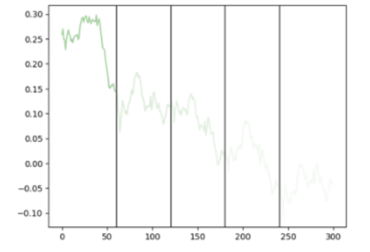
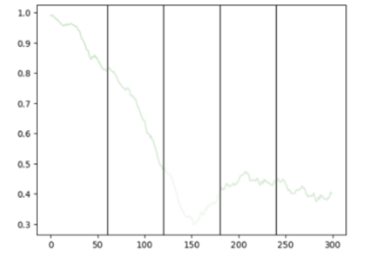
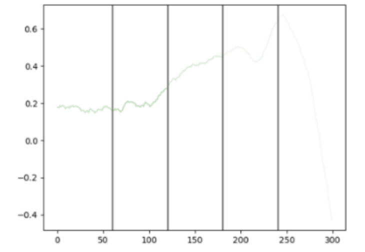
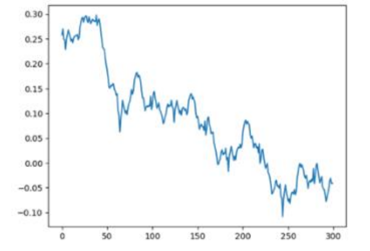
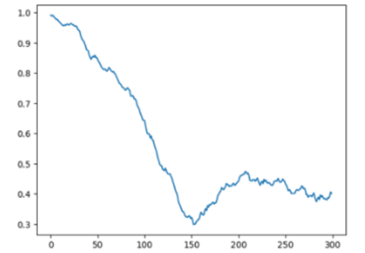
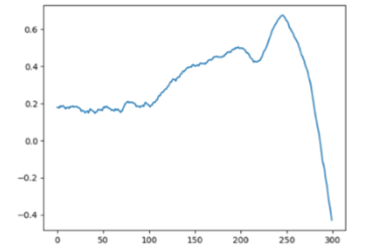
	Participant 1	Participant 2	Participant 3
2D Map			
			
X			
			
Y			
			

Table 6.22: Eye-tracking analysis for LA/LV

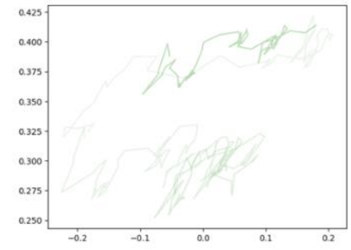
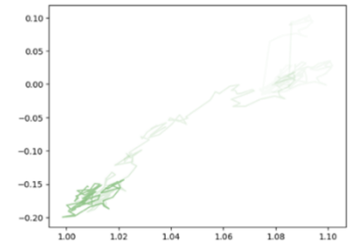
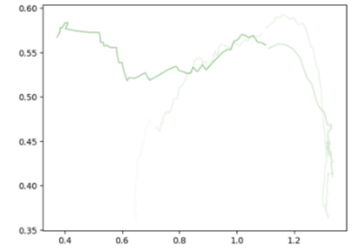
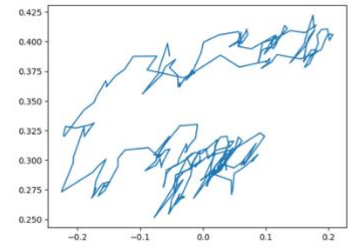
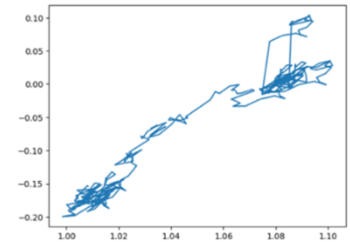
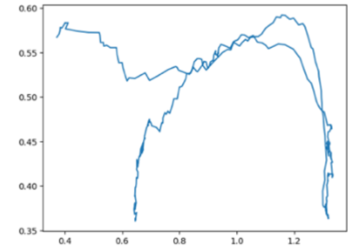
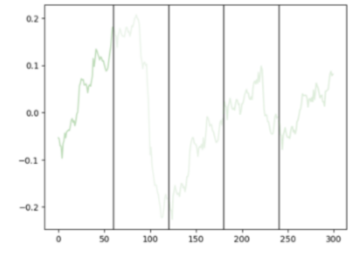
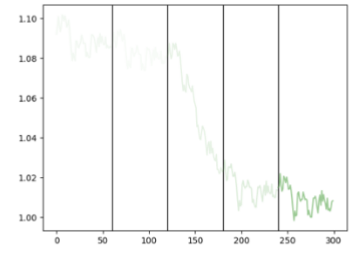
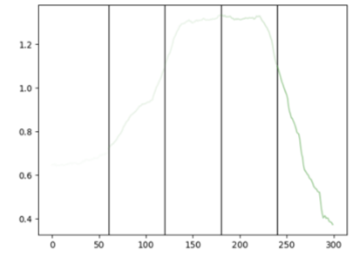
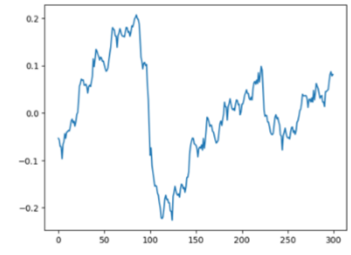
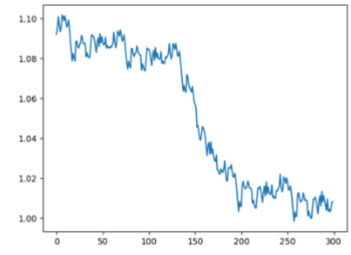
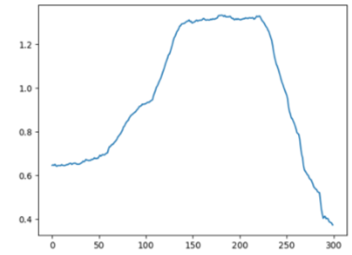
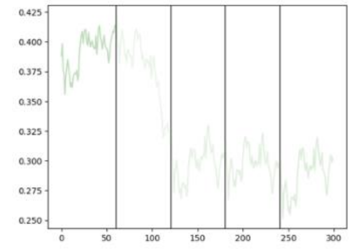
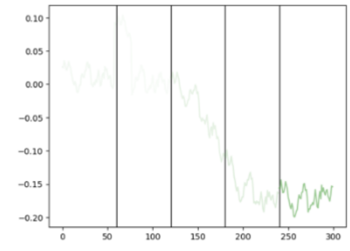
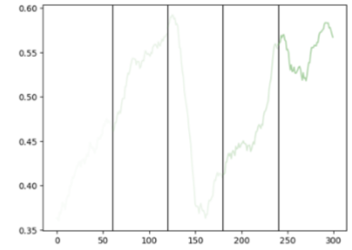
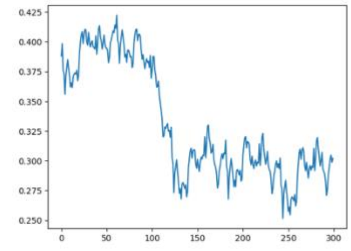
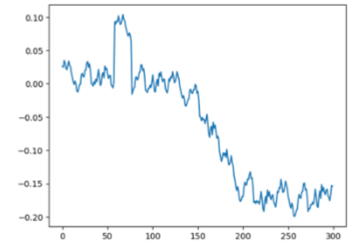
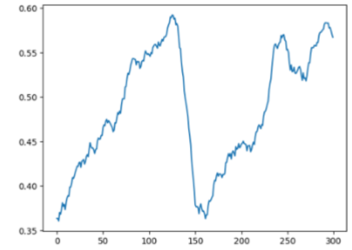
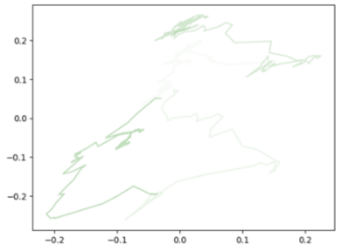
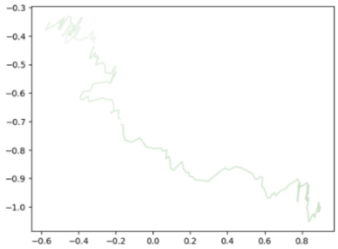
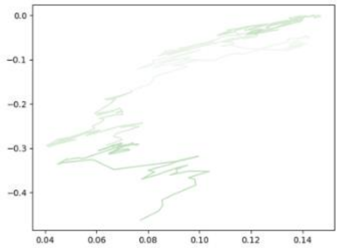
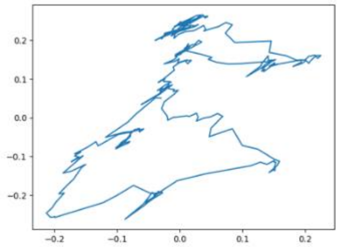
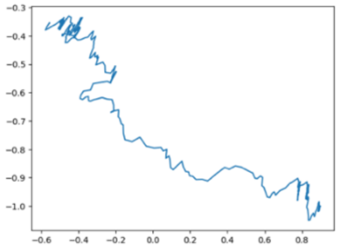
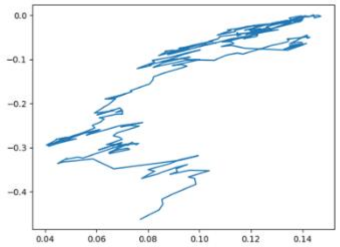
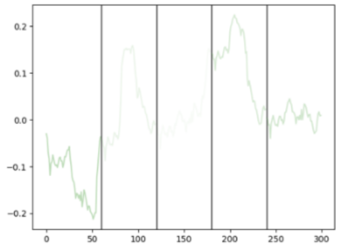
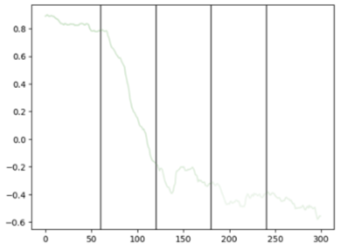
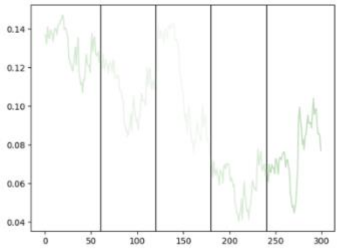
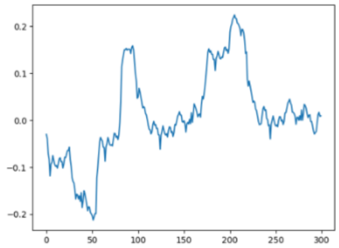
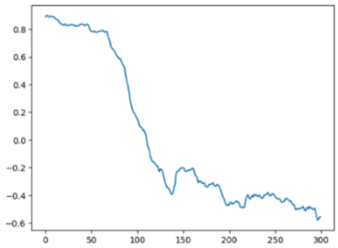
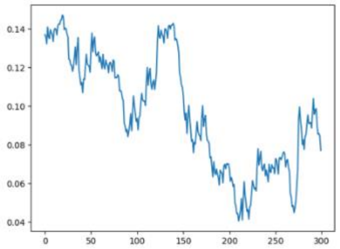
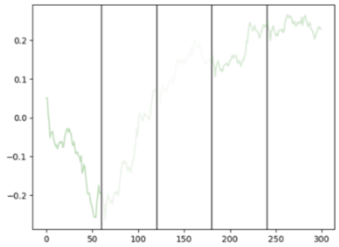
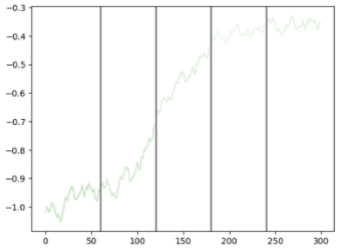
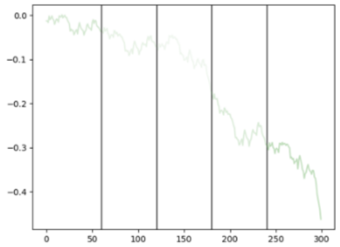
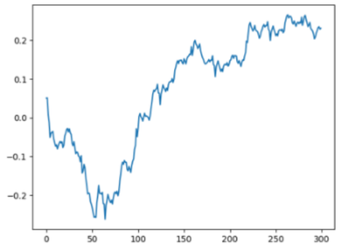
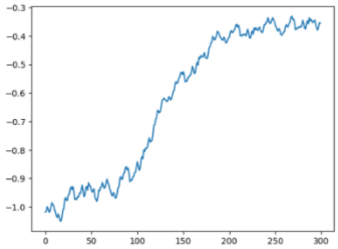
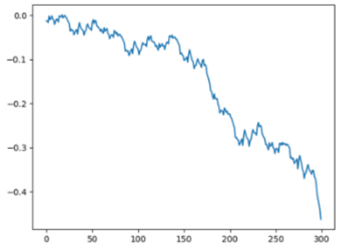
	Participant 1	Participant 2	Participant 3
2D Map			
			
X			
			
Y			
			

Table 6.23: Eye-tracking analysis for HA/LV

	Participant 1	Participant 2	Participant 3
2D Map			
			
X			
			
Y			
			

6.4 Discussion

6.4.1 Research Objective I: Investigate if Real-time Emotions of VR Users can be Detected Using Raw Time-series Physiological Signal Data.

In order to answer this research objective, we investigated different deep-learning architectures and their abilities to predict emotion based on three different window sizes (2.5 seconds, 5 seconds and 10 seconds). We found that a hybrid CNN-LSTM architecture outperformed individual CNN and LSTM models. The hybrid CNN-LSTM predicted the four classes corresponding to the four quadrants in the CMA with an accuracy of 63%. In order to compare our results to the literature and the results reported in Chapter 5, we provided accuracies of the complete data sample using a majority vote of the window predictions. Direct comparisons are hard to make due to the different pre-processing techniques, modalities and validation techniques. However, our deep learning approach outperforms the baseline results reported in Chapter 5 by 2.4% in four class classifications. These results indicate that emotions can be detected in real-time within virtual environments and that a deep learning approach that utilises raw time series data can outperform shallow methods. See section 6.3.1 for more detail.

6.4.2 Research Objective II: Investigate How can XAI be used to Provide Explanations for these Predictions.

In this research, we have presented a model-agnostic and time series modality-agnostic method of XAI. We have provided visualisations and insights for ECG, GSR and eye-tracking data using a single technique. Our XAI method utilises SHAP and the raw time series data broken down into segments in order to highlight the most important section of data to a given prediction (see 6.2.3 for more detail). This method allowed us to address a gap we identified in the literature: a lack of time series modality agnostic XAI implementations (Cortiñas-Lorenzo & Lacey, 2023) (see 2.8.3).

6.4.3 Research Objective III: Compare the XAI Outputs to the Feature Importance of the Baseline Model and Medical Literature to Further Understand the Inner Workings of the Model.

To address this research objective, we have presented global and local explanations and insights into this data. For the full findings, see section 6.3.2. Some of the key findings we found relate to both the comparison of the feature importance provided in Chapter 5 and the medical literature. When looking at comparing the results presented here to the Chapter 5 feature importance results, we found that ECG data played a greater role in this model and GSR had a lesser role compared to the results in Chapter 5 (see 6.3.2.1). Specifically, regarding GSR data, we found that the deep learning model doesn’t utilise peaks and valleys, whereas they were a main feature in Chapter 5. We hypothesised as to why this may be in section 6.3.2.2.2. Similar eye-tracking features were highlighted here when compared to Chapter 5. However, we did find that Higher frequency eye tracking data may provide improved results (see 6.3.2.2.3). ECG explanations provided some interesting insights, and features consistent with affective computing and medical literature, such as r peaks, were highlighted in our explanations. However, we also found that the t wave was also highlighted as an important feature of the deep learning model; this feature has not been extensively explored in affective computing or medical literature (see 6.3.2.2.1).

Chapter 7: Engagement Detection during Virtual Experiences and Investigation of EEG Signals and Their Feature and Electrode Importance

7.1 Introduction

Engagement is the level of involvement someone feels in a task (Ventura & Porfiri, 2020). The ability to detect and understand this has wide-ranging applications in various domains, including training, education, and healthcare. Engagement is a broad concept that refers to the "involvement, commitment, passion, enthusiasm, absorption, focused effort, zeal, dedication and energy" someone has for a task (Truss, 2014). Subjective methods such as questionnaires are generally used to measure engagement, and different questionnaires have been developed to measure engagement in different scenarios (Brockmyer et al., 2009). The role of engagement has been investigated across many domains and can play a prominent role throughout various areas of healthcare; for example, there is a strong relationship between patient engagement in treatment and the outcome of psychotherapy (Horvath et al., 2011). With this being the case, a successful at-home therapy system would need to recognise and react to engagement to provide the best treatment. See section 2.4 for further discussion on the role of engagement during mental health therapies.

Statistical and ML analysis have been used to investigate engagement, but creating labelled data and eliciting engagement can be challenging. Video games are the most common stimuli to evoke carrying levels of engagement (Gábana Arellano et al., 2016; Chanel et al., 2011). Studies have used various ways to elicit engagement; some successful studies have been based around making participants play the same game but with a different level of competitiveness, i.e., playing a game solo, collaboratively with someone and then competitively with someone (Gábana Arellano et al., 2016). Game difficulty has also been used successfully to evoke engagement; the thought is that if a user is bored if a game is too easy, anxious if a game is too hard and engaged when the difficulty is just right (Chanel et al., 2011; Chanel et al., 2008).

Various physiological signals have been used to detect engagement, one of the most common being Electroencephalogram (EEG) signals. EEG signals correlate with engagement; Rogers et al., (2020) measured engagement in virtual reality therapy using a single left pre-frontal electrode. This paper concludes that frontal theta power in healthy adults provides a valid measure of user engagement within virtual reality (VR). Another study uses frontal EEG for engagement classification for affective cinema (Abadi et al., 2013). The individual contributions of GSR, EEG and facial tracking were investigated. It was reported that each modality significantly encodes the participants' engagement level across a range of video clips. See section 2.4.1 for more discussion on physiological responses to engagement and section 2.6.2 for further discussion on engagement recognition.

The ability to detect engagement within a virtual environment has implications for healthcare, specifically immersive virtual therapy. Patient engagement in therapy is strongly linked to treatment outcome (Gaston, 1998; Gomes-Schwartz, 1978; Zuroff et al., 2016); therefore, the ability to monitor engagement could help clinicians while conducting virtual immersive therapy. There is, however, limited research into detecting engagement within virtual environments. The use of EEG signals is a practical decision when considering immersive virtual therapy because it is reasonable to think that EEG sensors could be built into a headset without any separate devices in the future. These VR headsets are already starting to be developed and sold mainly for research purposes (Ag, 2023).

To this end, in this study, we investigate and present an engagement detection machine learning (ML) model that uses EEG signals. These signals were recorded while participants undertook low, moderate, and highly engaging tasks in VR. With VR becoming increasingly more used as a tool within therapy (Dellazizzo et al., 2020) and the benefits that VR brings, such as the increased presence and immersion stated and proven in previous chapters, VR was the best choice of stimuli for this research. We further investigate the importance of electrodes and specific features to engagement. Our research objectives for this were as follows:

i) Understand whether VR is a reliable stimulus for engagement recognition. This relates to overall thesis objective 2 (see 1.2)
ii) Investigate if engagement can be detected within virtual environments using ML and compare the results to literature and studies that utilise non-immersive stimuli (see 1.2)
iii) Investigate the specific features of data that indicate engagement and compare these to affective computing and medical literature. This relates to overall thesis objective 5 (see 1.2)

7.2 Methodology

7.2.1 Data Collection

7.2.1.1 Participants

Eighteen individuals (male=10, female=8, average age = 22.82(SD=8.52)) volunteered to participate in this study. Ethical approval for the study was obtained from the Central Research Advisory Group at the University of Kent. To be eligible to partake, participants had to be healthy adults and must not have been:

- Blind or colour-blind
- Unable to understand verbal English instruction
- aged under 18
- Suffer from motion sickness when using VR

7.2.1.2 Physiological Measures

In our study, we collected EEG data. EEG is a recording of brain activity that aids in diagnosing and monitoring various conditions that impact the brain (NHS 2022; da Silva, 2009). Numerous research papers have delved into the relationship between EEG data and engagement levels, offering frameworks and highlighting the importance of specific electrodes and features (Hafeez et al., 2021; Ruqeyya et al., 2022; Berka et al., 2007). To record EEG signals, we used the Biosemi system, which included the ActiView software. Various cap sizes were used, with 32 channels (Biosemi 2001).

It is worth noting that while many studies combine EEG with other modalities, our approach focuses solely on EEG recordings. Our rationale for this choice is to establish a minimalistic system with potential future integration into VR headsets.

7.2.1.3 Questionnaire

The Game Engagement Questionnaire (GEQ), developed by Brockmyer et al., is a cost-effective and efficient tool for assessing the engagement levels of video game players across four dimensions: immersion, presence, flow, and absorption. Despite its strengths, some limitations have been noted in its application (Huynh et al., 2018). Gamers may find it challenging to accurately recall their gaming experiences once they have finished playing, often leading to scores reflecting their end-of-session feelings rather than an overall engagement assessment. Fortunately, our study's relatively short game sessions, each lasting 10 minutes, mitigate the impact of these limitations.

In our research, we intend to utilise the GEQ scores to validate our stimuli and ensure they evoke the desired response. This questionnaire has also been employed in similar studies (Psaltis et al., 2018), which aimed to leverage ML techniques to recognise engagement during gameplay. Notably, this related study employed a slightly modified version of the GEQ to annotate their data, further demonstrating its adaptability and utility in assessing engagement levels in gaming contexts.

7.2.1.4 VR Headset

For our study, we used the Meta Quest 2 VR headset (Meta Quest 2: Immersive all-in-one VR headset: Meta store 2023) as our VR platform of choice. This headset was selected based on several reasons: The Meta Quest 2 is an all-in-one VR headset, eliminating the need for additional external hardware or wires. The ease of adjustment ensures a comfortable fit for various users. The headset is designed to accommodate users who wear glasses, making it inclusive and accessible to a broader range of participants. It has sufficient computational power to run our study's specific games and videos. Notably, the strap placement and fit of the headset allow us to use EEG sensors while the user is wearing the headset.

7.2.1.5 Stimuli

We decided to use a video game as engagement stimuli as video games have been used in similar studies investigating and eliciting various levels of engagement (Monteiro et al., 2018; Chanel et al., 2008). In order to stimulate various levels of engagement, we use various participation levels and difficulties in a game. Emotion and flow theories state that strong involvement in a task occurs when the difficulty of a task is equal to the skill level of the individual. Too much challenge would raise anxiety, and not enough would induce boredom (Chanel et al., 2008). The stimuli we used to elicit the three levels of engagement were watching gameplay (low engagement), playing on a difficulty that is too hard (moderate engagement) and playing at the participant's selected difficulty (high engagement).

The chosen game was Tetris (Tetris® effect: Connected 2023)(Figure 7.1). Tetris was chosen because it is a cognitively challenging game with simple rules and controls. The difficulty is also adjustable to cater for participants with a range of skill levels and experiences with video games. Tetris has also been used in other studies with similar procedures to elicit engagement (Chanel et al., 2008). In particular, the marathon mode of Tetris will be used; this is an endless version where the player continues until failure.

We have used the commercially available Tetris Effect: Connected (VR Oculus) (Tetris® effect: Connected 2023) for the medium and high engagement sections and a YouTube video of someone playing the Tetris effect (Robinson, 2018) for the low engagement section. Screenshots can be found in Figure 7.1.

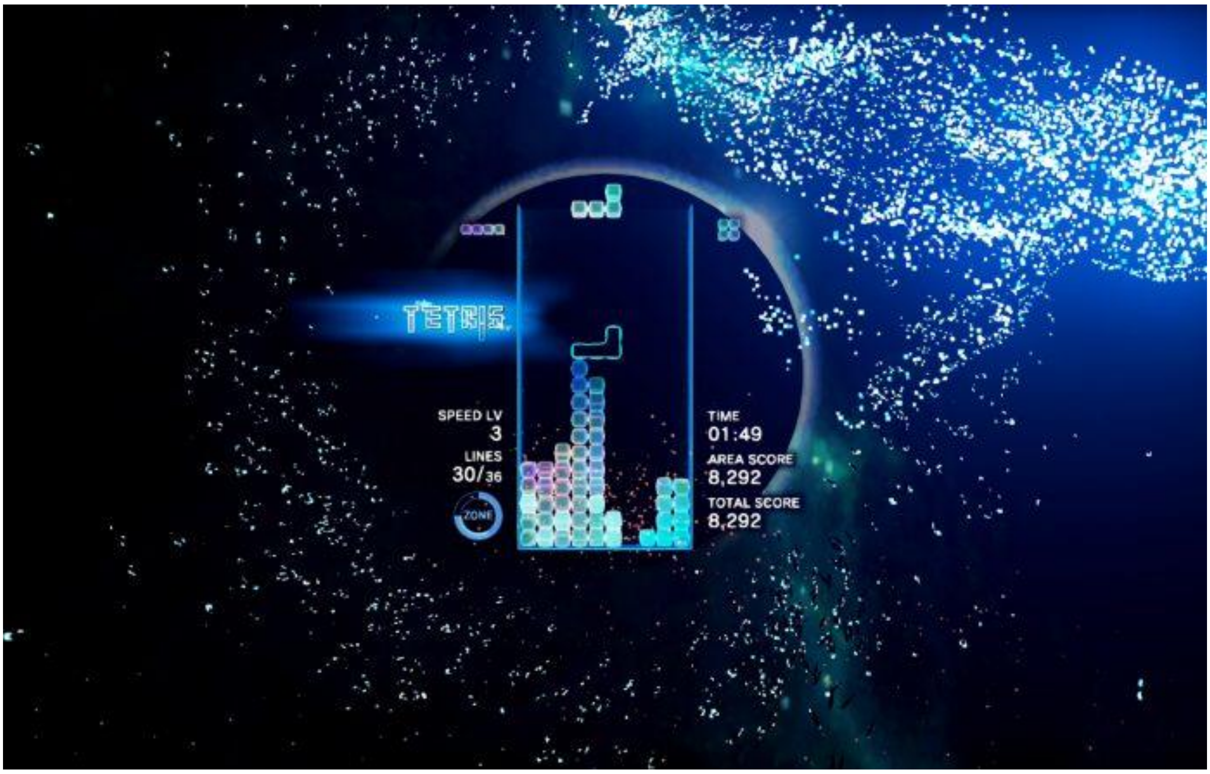


Figure 7.1: Tetris Effect: Connected gameplay screenshot

7.2.1.6 Experimental Procedure

In our study, we followed a procedure closely resembling the one outlined by Chanel et al. (2008), which has been proven effective in eliciting varying levels of engagement. Each participant made a single visit to our laboratory, with each session lasting approximately two hours. We ensured consistency in instructions by implementing a verbal instructions protocol for all participants.

1. Initially, participants were briefed on the procedure, explained Tetris rules, and had sensors applied.
2. Our study conductors then carefully adjusted the headset and straps to ensure proper electrode placement without interference.
3. After ensuring signal accuracy and noise-free data, participants could freely play Tetris, familiarise themselves with the controls, and choose their preferred difficulty level. This phase marked the preparatory stage of our study.

Following the preparatory phase, we transitioned into the data collection/recording phase. Participants underwent three distinct phases of the procedure in a randomised order, each involving 10 minutes of gameplay or observation of the Tetris VR version, followed by an engagement questionnaire (as detailed in section 6.2.1.3). To induce different levels of engagement, we selected three unique tasks, all centred around Tetris in VR (as described in section 6.2.1.5). The three different stages were as follows:

i) Low engagement, where participants watched Tetris gameplay.

ii) High engagement, involving active gameplay at their preferred difficulty level (selected by the participant in the preparatory stage).

iii) Moderate engagement, where participants play Tetris at a deliberately challenging difficulty (four difficulties higher than the participants preferred difficulty).

Upon completion of these phases, we removed the headset and sensors and conducted a debriefing session with the participants.

7.2.2 Data Processing and Analysis

7.2.2.1 Pre-Processing

The dataset consists of data collected from 18 participants, each completing three trials, resulting in a total of 54 complete samples. These samples were further segmented into 16,200 five-second data segments. The dataset is fully balanced across three classes, representing low, moderate, and highly engaging data points. Specifically, there are 18 data points per class in the complete samples and 5,400 data points per class in the segmented data, ensuring equal representation for each engagement level across the dataset.

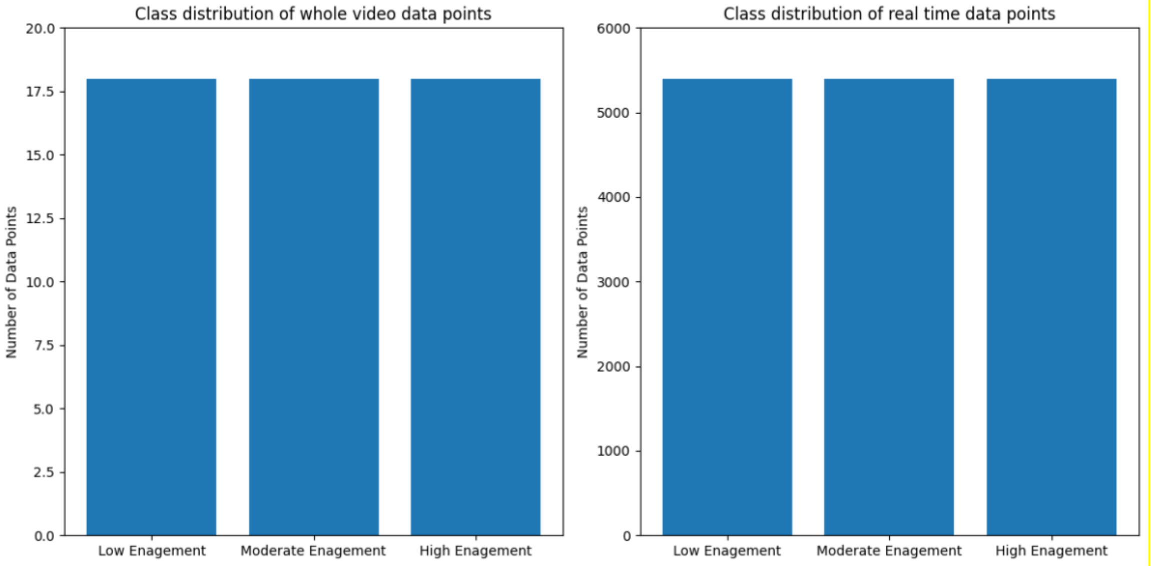


Figure 7.2: Class Distribution

We followed Luck's (2014) recommended procedure to pre-process the EEG signals. The first step was to resample from 512hz to 250hz, which was done for computational efficiency. Using the Nyquist limit (Landau, 1967), the highest frequency we are looking at is 60HZ, so the data needed to be sampled at least 125Hz to see those signals. The rest of the pre-processing was done using the MNE Python library (Gramfort, 2013). The following steps were taken:

- The signal was re-referenced to the average of all electrodes.
- Then, it was notch-filtered at 50Hz to remove power line noise
- Then the signal was bandpass filtered between 1 and 60Hz
- Independent Component Analysis (ICA): The signals were then processed using ICA to remove artefacts such as eye blinks, muscle movements, and other non-neural noise.
    - The Infomax ICA algorithm separates mixed signals into independent components by finding a way to "unmix" them. It works by maximising the statistical independence of the components using an iterative process. For EEG, this helps isolate brain signals from artefacts like eye blinks or muscle movements. The algorithm adjusts the unmixing matrix step by step to achieve the best separation of independent sources. Here, we used the infomax ICA algorithm with 20 components and a maximum of 50,000 iterations.
    - After extracting the components, artefacts were identified and rejected using MNE’s ICLabel (Li et al., 2022). ICLabel is an ML classifier that assigns a probability to each component, indicating its likelihood of being brain activity or non-brain artefacts.

Components classified as non-brain artefacts with a probability of 80% or higher were removed from the signal. The artefacts classified are as follows:

    - Brain
    - Muscle
    - Eye blink
    - Heart beat
    - Line noise
    - Channel noise
    - Other

- Files were saved to disk in MNE format for analysis

We also used a sliding window with a size of 5 seconds and a step of 1 second to follow the same values identified in Chapter 6. This was to make the data suitable for the real-time approach.

7.2.2.2 Feature Extraction

In our initial pilot testing, we extracted frequency and time domain features. The frequency domain features we extracted were the logarithms of the Power Spectral Density from delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (14-30 Hz), and gamma (> 30 Hz) bands from all channels. This was done using the MNE feature extraction library (Schiratti et al., 2018). These features have commonly been used throughout affective computing literature (Aspiras & Asari, 2011; Soleymani et al., 2014).

Time domain statistical features such as mean, standard deviation, kurtosis, etc., have been used with success in various EEG classification tasks, including emotion detection, utilising the DEAP dataset (Feradov and Ganchev, 2015) and in classifying sleep stages (Diykh et al., 2016). In addition, they are computationally efficient, which is an important consideration if used in a real-time detection system. We extracted each channel's mean, standard deviation, variance, skew and kurtosis using the MNE feature extraction library (Schiratti et al., 2018). This produced a total of 160 features. The features and their equations can be found in Table 7.1.

Table 7.1: Time Domain EEG Features

Feature	Equation
Mean - the sum of all values in a dataset divided by the total number of values, representing the central tendency of the data.	$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$
Standard Deviation - measures the average amount by which individual values in a dataset differ from the mean, providing an indication of the spread or variability of the data.	$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$
Variance - quantifies the average squared difference between each value in a dataset and the mean, serving as a measure of how spread out the values are.	$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$
Skew - measures the asymmetry of the distribution of a dataset; positive skew indicates a tail on the right, negative skew indicates a tail on the left, and a skew of zero indicates a symmetric distribution.	$\text{Skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^3}{\sigma^3}$
Kurtosis - measures the "tailedness" of a dataset's distribution; high kurtosis indicates heavy tails and sharp peaks, while low kurtosis suggests light tails and a flatter distribution compared to a normal distribution.	$\text{Kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^4}{\sigma^4}$

7.2.3 Model Development and Analysis

7.2.3.1 Machine Learning

We initially referred to the literature when selecting the initial models to utilise in this study. A review was conducted (Wang & Wang, 2021) that looked at emotional feature extraction and classification using EEG signals. This study stated that some of the most popular models were support vector machines (SVM) and random forests (RF). In addition to these models, we also included some ensemble methods, such as extra trees classifier (ET) and gradient boosting (GB). We included these ensemble methods because they commonly seem to outperform non-ensemble methods. Feature importance is then carried out (see section 6.2.3.2 for more detail on the feature importance method). All results reported have been obtained using nested leave one participant out methodology.

Pilot testing was initially carried out with the all-feature sets and models. We used the binary classification task for the initial pilot testing. The results for the highest performing model for each feature set can be found in Tables 7.2 and 7.3. From this initial testing, we found that there was not a particular model that consistently outperformed all other models, but the Time domain feature set did seem to outperform the other feature sets consistently. Therefore, we proceeded to the main testing with all models and the time domain feature set. From here forward, the complete feature set refers to the complete time domain feature set.

Table 7.2: Binary Whole Video Pilot Results

Feature set	Precision	Recall	F1	Accuracy
Time Domain (GB)	**0.75 (SD=0.46)**	0.64 (SD=0.50)	**0.69 (SD=0.46)**	**0.71 (SD=0.26)**
Frequency Domain (RF)	0.69 (SD=0.41)	**0.68 (SD=0.32)**	0.67 (SD=0.37)	0.67 (SD=0.32)
Combined (RF)	0.61 (SD=0.41)	0.61 (SD=0.35)	0.61 (SD=0.39)	0.61 (SD=0.35)

Table 7.3: Binary Sliding-Window Pilot Results

Feature set	Precision	Recall	F1	Accuracy
Time Domain (ET)	0.565 (SD=0.101)	**0.571 (SD=0.241)**	**0.568 (SD=0.114)**	**0.566 (SD=0.060)**
Frequency Domain (SVM)	0.565 (0.163)	0.565 (SD=0.146)	0.564(SD=0.167)	0.565(SD=0.146)
Combined (RF)	**0.566 (SD=0.176)**	0.566 (SD=0.154)	0.565 (SD=0.170)	0.565 (SD=0.150)

For each classification task (see below for info on the four classification tasks), each model was initially tested utilising the complete feature set. The selected features, resulting from the feature importance analysis, were then used with the highest-performing model. Leave-one-participant-out validation was used for both the complete feature set and the selected feature set, and the average and standard deviation of performance metrics were presented. Four different classification tasks were conducted:

i) Binary classification of the complete sample: classification of high and low engagement using the complete sample to extract features

ii) Three class classifications of the complete sample: classification of high, medium and low engagement using the complete sample to extract features

iii) Binary classification of sliding windows: classification of high and low engagement using a sliding window with a size of 5 seconds and a step of 1 second

iv) Three class classifications of sliding windows: classification of high, medium and low engagement using a sliding window with a size of 5 seconds and a step of 1 second

We conducted binary and three-class classifications to make our results more comparable with baseline results reported in the literature. This also allows us to investigate classifying vastly different engagement levels (low and high) and whether slight differences in engagement can be detected (low, medium, and high).

7.2.3.2 Feature Importance

We looked at two sets of importance in this study: feature and electrode. The reasons for carrying out feature importance were to enable the development of more efficient and accurate ML models and compare results to affective computing and medical literature. Any feature with a performance score greater than 0.01 was added to the selected feature set. A threshold of 0.01 left a sensible number of features for each classification task, between 29 and 38 (roughly 20% of the complete feature set). A feature importance threshold was used to select features instead of a specific number of features to ensure no impactful features were left out. The reason for the importance of electrodes was to allow comparison to EEG-specific literature and inform the development of integrated VR EEG sensors.

To determine feature importance, we used the same techniques as in Chapter 3 (see 3.2.3.4), as we predominantly use tree-based models here. Feature importance was carried out for each classification task separately; therefore, we could see if certain electrodes/features were more important when classifying three classes than two. In our results, we present the ranking of all the electrodes and all features with an importance score greater than 0.01.

7.3 Results

7.3.1 Validation of Dataset

We used the GEQ questionnaire (see more info in section 6.2.1.3) and followed the procedure outlined in (Psaltis et al., 2018) to calculate the engagement scores. With these results, we wanted to verify that our stimulus evoked the correct levels of engagement and ensure that the results and the rated levels of engagement were significantly different. In order to compare the questionnaire results, we performed a one-way ANOVA test.

The analysis we performed verifies the stimulus-evoked the intended levels of engagement. We observed a substantial contrast between the low engagement and highly engaging stimuli, while the disparity between moderately engaging and highly engaging stimuli was comparatively smaller. Interestingly, we found that the high and low engagement populations did not overlap within one standard deviation. However, the moderate engagement level overlaps with high and low categories, indicating that its classification poses a greater challenge. Our findings were further validated through a one-way ANOVA test, confirming the significant difference among the stimuli and their corresponding engagement ratings ($p<0.05$).

Table 7.4: Rated Levels of Engagement

Intended Level of Engagement	Mean Rated Engagement	Standard Deviation Rated Engagement
Low	-10.64	9.02
Moderate	4.02	13.72
High	10.58	9.43

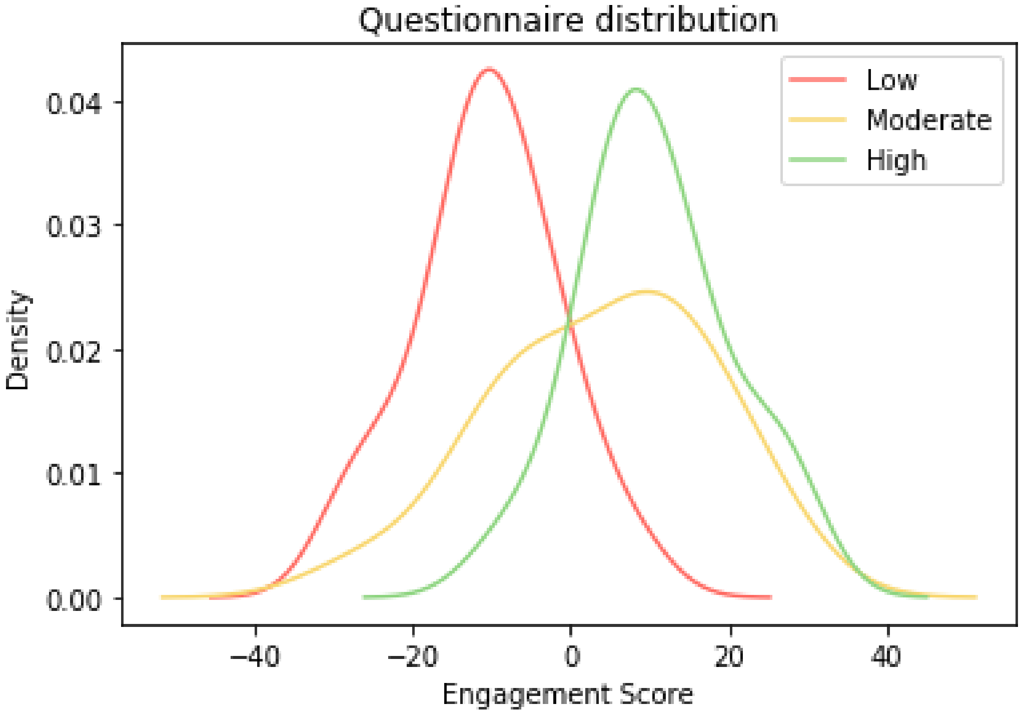


Figure 7.3: Engagement questionnaire response distribution

7.3.2 Machine Learning Results

7.3.2.1 Binary Classification of Complete Sample

For this task, the chosen model was a GB model. Initially, when utilising the complete set of features, the model exhibited an accuracy of 0.71 (SD=0.26). However, a refinement in feature selection led to an enhancement, resulting in an improved accuracy of 0.75 (SD=0.26), as shown in Table 7.5. Out of the pool of features, 38 were selected based on their importance scores (importance >0.01), these features can be found in Figure 7.6. These selected features improved the model's performance, particularly in terms of recall, minimising instances where engaged samples were misclassified as not engaged.

The top statistical features that emerged were standard deviation, kurtosis, and skewness. This implies that the distribution and variation of the data were crucial for the classifications. The impact of electrode placements was also investigated, revealing that two specific placements, FC6 and C3, substantially influenced the model's performance. This comprehensive analysis and fine-tuning of features optimised the model's performance and highlighted the pivotal role of certain statistical features and electrode placements.

Table 7.5: Binary Classification Results

	Precision	**Recall**	**F1**	**Accuracy**
SVM Complete Feature Set	0.45 (SD=0.25)	0.49 (SD=0.39)	0.47 (SD=0.32)	0.50 (SD=0.20)
RF Complete Feature Set	0.49 (SD=0.38)	0.52 (SD=0.28)	0.50 (SD=0.33)	0.50 (SD=0.28)
GB Complete Feature Set	0.75 (SD=0.46)	0.64 (SD=0.50)	0.69 (SD=0.46)	0.71 (SD=0.26)
ET Complete Feature Set	0.53 (SD=0.33)	0.56 (SD=0.25)	0.54 (SD=0.26)	0.54 (SD=0.24)
GB Selected Feature Set	0.77 (SD=0.45)	0.71 (SD=0.47)	0.74 (SD=0.44)	0.75 (SD=0.26)

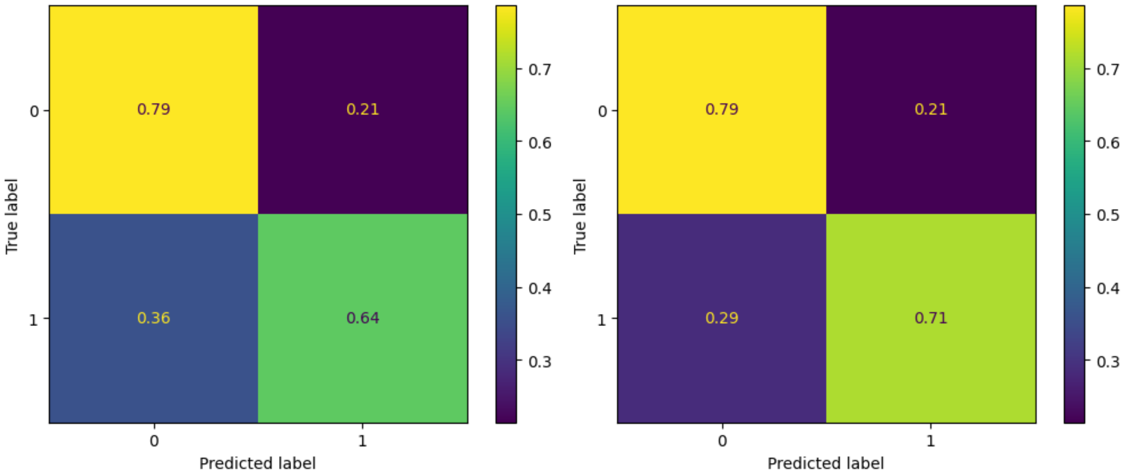


Figure 7.4: Confusion matrix for binary complete feature set (left) and selected feature set (right)

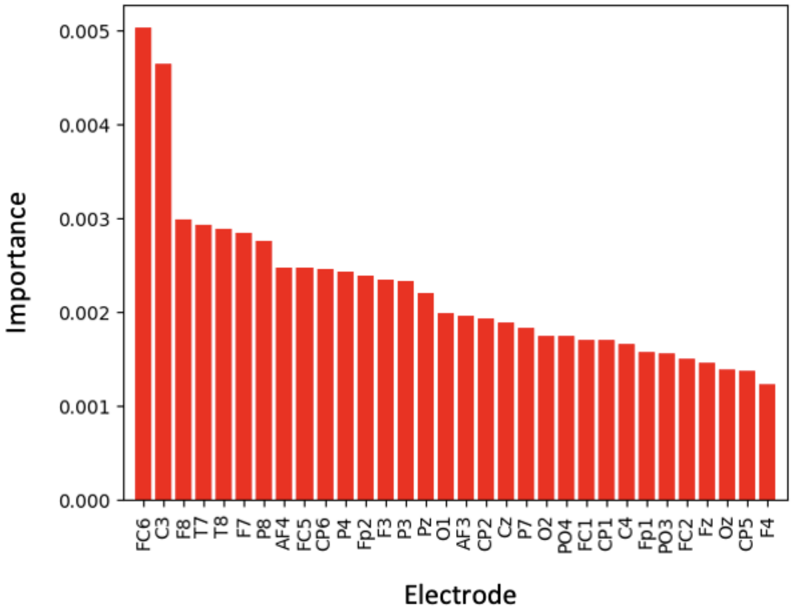


Figure 7.5: Binary classification whole video electrode importance's

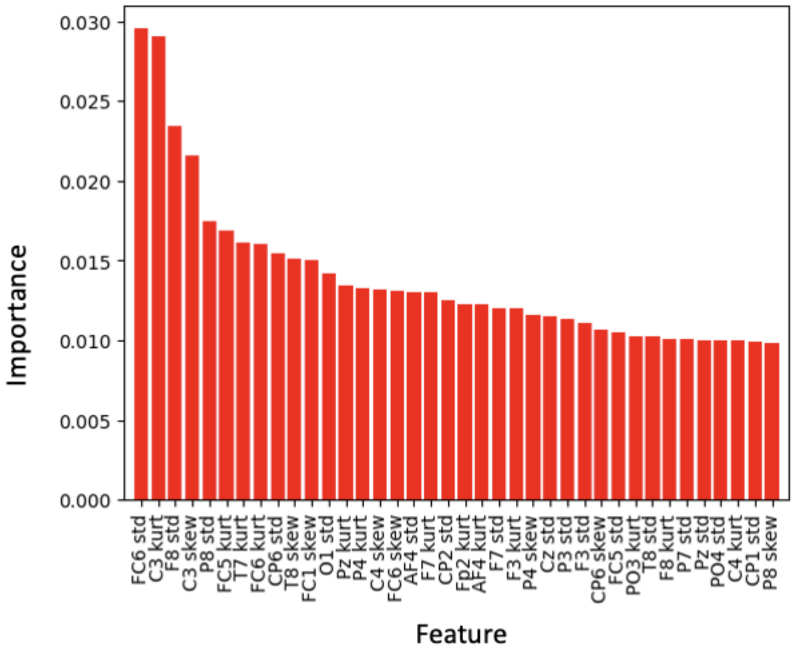


Figure 7.6: Binary classification whole video feature importance's (>0.01)

7.3.2.2 Binary Classification on Sliding Window

The ET classifier was the chosen model in this task, initially achieving an accuracy of 0.566 (SD=0.060). During the feature selection process, 29 features were selected based on their importance scores (>0.01). When this feature set was utilised, it improved the accuracy of the ET, reaching 0.610 (SD=0.173).

The impact of this selected feature set on recall differed from the findings presented in section 7.3.2.1. Contrary to the previous results, there was a reduction in recall but a simultaneous increase in precision. This meant that when the model predicted engagement, there was a higher likelihood that the prediction was correct, potentially instilling greater confidence in the system's predictions despite the decrease in overall recall.

Examining the specific features, three electrodes, FC6, F8, and F7, emerged as particularly influential in the classification process. Additionally, it is worth noting that standard deviation was the only statistical feature chosen among the selected features.

Table 7.6: Binary Classification Real-Time Results

	Precision	**Recall**	**F1**	**Accuracy**
SVM Complete Feature Set	0.492 (SD=0.198)	0.460 (SD=0.300)	0.475 (SD=0.205)	0.487 (SD=0.125)
RF Complete Feature Set	0.572 (SD=0.178)	0.522 (SD=0.303)	0.546 (SD=0.200)	0.566 (SD=0.131)
GB Complete Feature Set	0.549 (SD=0.208)	0.483 (SD=0.372)	0.514 (SD=0.252)	0.543 (SD=0.154)
ET Complete Feature Set	0.565 (SD=0.101)	0.571 (SD=0.241)	0.568 (SD=0.114)	0.566 (SD=0.060)
ET Selected Feature Set	0.637 (SD=0.234)	0.511 (SD=0.387)	0.567 (SD=0.285)	0.610 (SD=0.173)

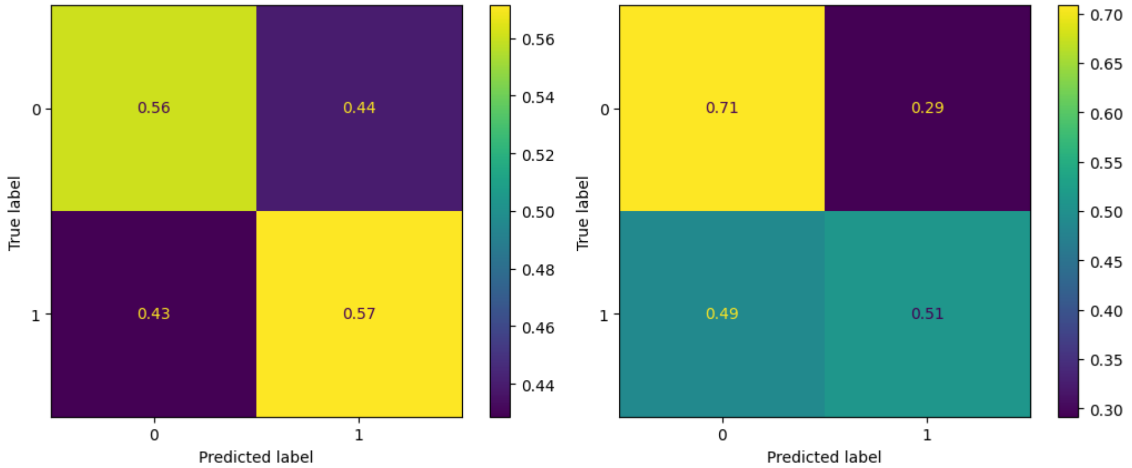


Figure 7.7: Confusion matrix for real-time binary classification using the complete feature set (left) and the selected feature set (right)

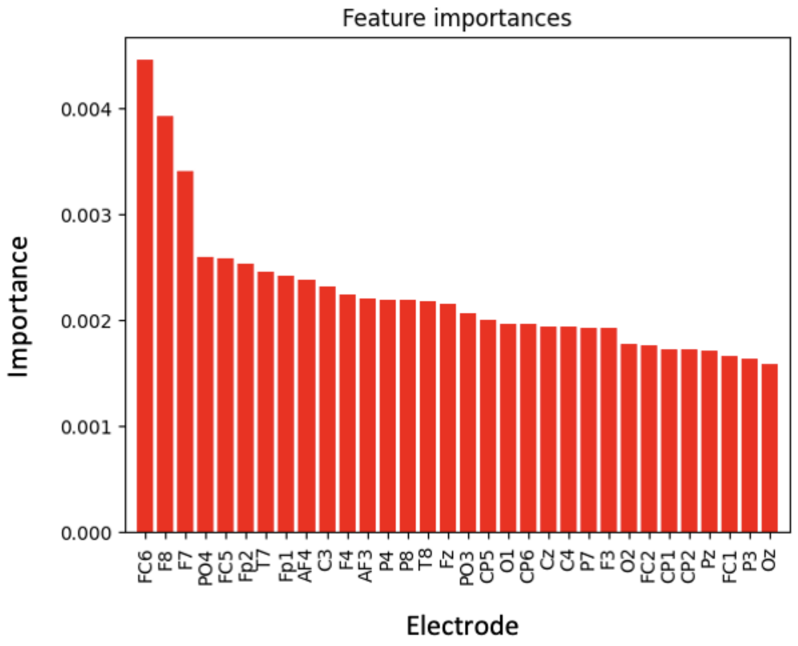


Figure 7.8: Binary classification whole video Electrode importance's

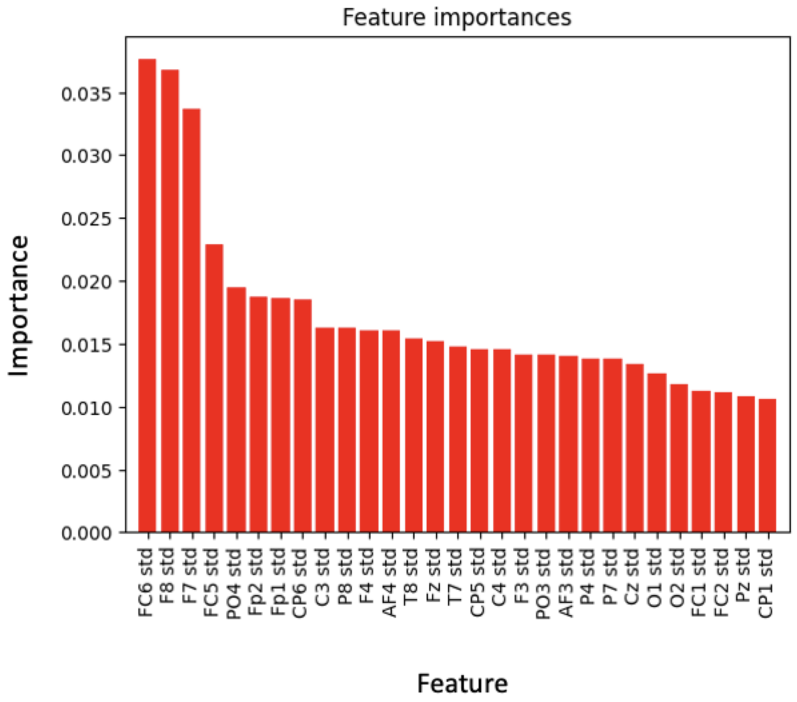


Figure 7.9: Binary classification whole video feature importance's (>0.01)

7.3.2.3 Three-Class Classification of Complete Sample

As expected, the three-class classification was a much more complex classification task than the previous tasks, in particular, the prediction of the moderately engaging task (label 1). This possible problem was identified when analysing the reported engagement levels (section 6.3.1). In this analysis, an ET model was employed, initially yielding an accuracy of 0.381 (SD=0.221). After feature selection, 32 features were chosen based on their importance, surpassing the threshold of 0.01. This selected feature set improved precision, recall, F1 score, and overall accuracy. Upon using the selected feature set, the accuracy notably increased to 0.524 (SD=0.171), signifying a substantial enhancement in the model's performance.

In assessing the importance of electrodes, C3 emerged as the most crucial electrode, followed closely by FC6, as indicated in Figure 7.11. Similarly to the previous tasks, the selected features were limited to standard deviation, kurtosis, and skewness.

Table 7.7: Three Class Classification Results

	Precision	**Recall**	**F1**	**Accuracy**
SVM Complete Feature Set	0.330 (SD=0.117)	0.310 (SD=0.158)	0.231 (SD=0.127)	0.310 (SD=0.158)
RF Complete Feature Set	0.265 (SD=0.086)	0.262 (SD=0.142)	0.258 (SD=0.095)	0.262 (SD=0.142)
GB Complete Feature Set	0.261 (SD=0.175)	0.262 (SD=0.233)	0.261 (SD=0.191)	0.262 (SD=0.233)
ET Complete Feature Set	0.367 (SD=0.187)	0.381 (SD=0.221)	0.369 (SD=0.197)	0.381 (SD=0.221)
ET Selected Feature Set	0.514 (SD=0.174)	0.524 (SD=0.171)	0.508 (SD=0.175)	0.524 (SD=0.171)

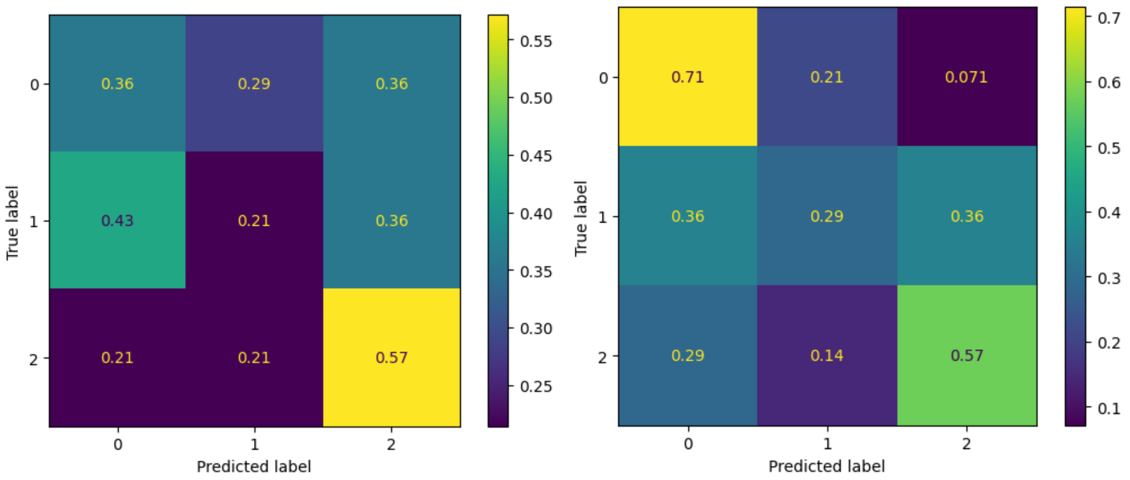


Figure 7.10: Confusion matrix for three class classifications using the complete feature set (left) and the selected feature set (right)

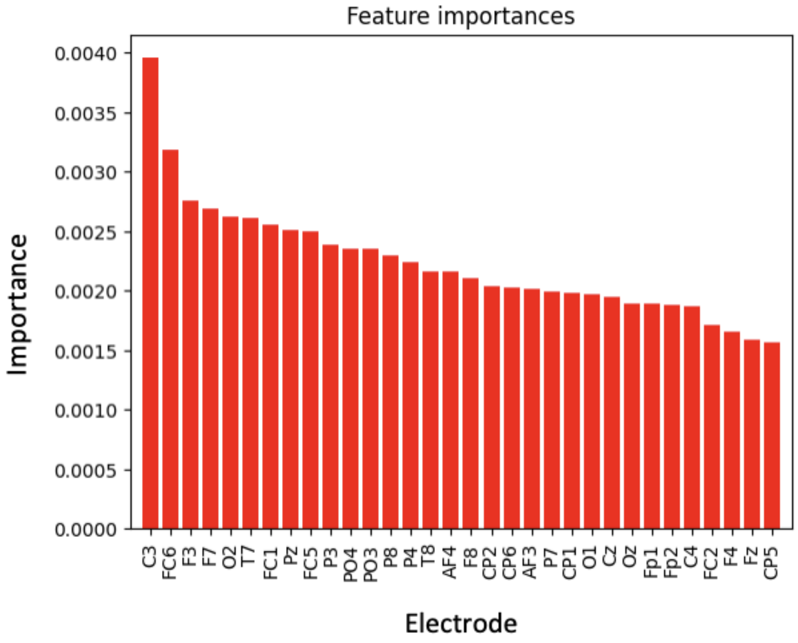


Figure 7.11: Three class classification whole video electrode importance’s

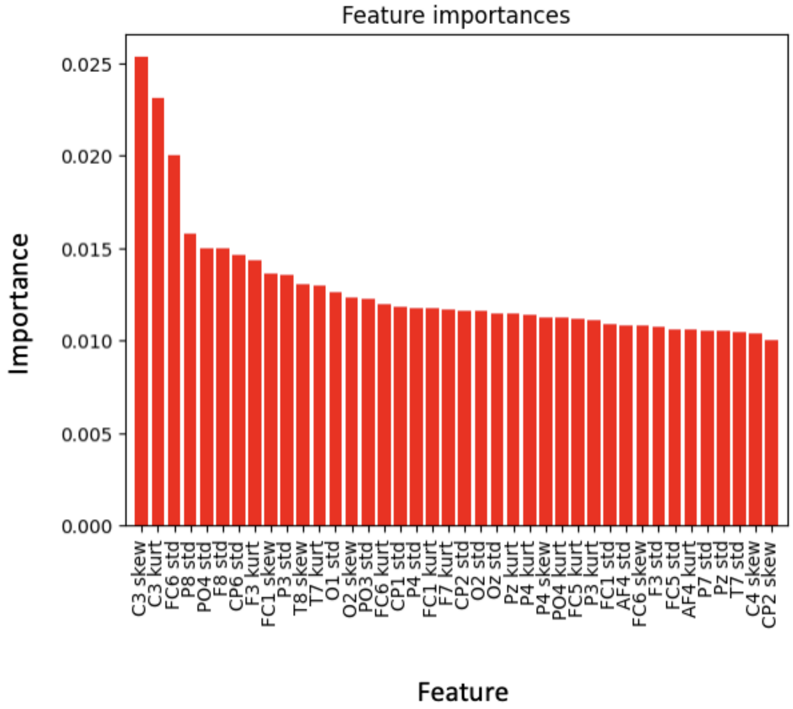


Figure 7.12: Three class classification whole video feature importance's (>0.01)

7.3.2.4 Three-Class Classification of Sliding Windows

This task utilised an ET model; the initial accuracy was 0.385 (SD=0.098). Like the previous task, moderately engaging (label 1) was the most challenging class to classify. During the feature selection process, 31 features were selected. This task marked the first instance of the four where the model's performance declined when the selected feature set was employed. The selected features decreased accuracy, dropping to 0.376 (SD=0.079).

In contrast to previous tasks, the importance values in this case were more evenly distributed among the features. F8, FC6, and F7 emerged as the electrodes with the highest importance values. Similarly to the other real-time classification tasks, only standard deviation statistical features were chosen, highlighting the consistent significance of this analysis.

Table 7.8: Three Class Classification Real-time Results

	Precision	**Recall**	**F1**	**Accuracy**
SVM Complete Feature Set	0.313 (SD=0.084)	0.329 (SD=0.069)	0.288 (SD=0.066)	0.329 (SD=0.069)
RF Complete Feature Set	0.375 (SD=0.072)	0.377 (SD=0.064)	0.374 (SD=0.059)	0.377 (SD=0.064)
GB Complete Feature Set	0.365 (SD=0.086)	0.366 (SD=0.091)	0.365 (SD=0.078)	0.366 (SD=0.091)
ET Complete Feature Set	0.382 (SD=0.092)	0.385 (SD=0.098)	0.382 (SD=0.099)	0.385 (SD=0.098)
ET Selected Feature Set	0.375 (SD=0.132)	0.376 (SD=0.077)	0.371 (SD=0.080)	0.376 (SD=0.079)

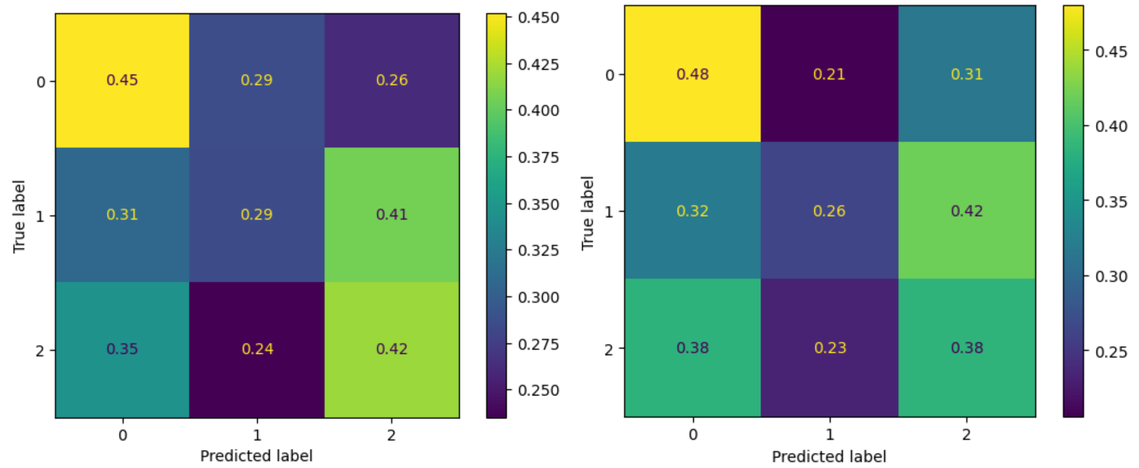


Figure 7.13: Confusion matrix for three class classifications using the complete feature set (left) and the selected feature set (right)

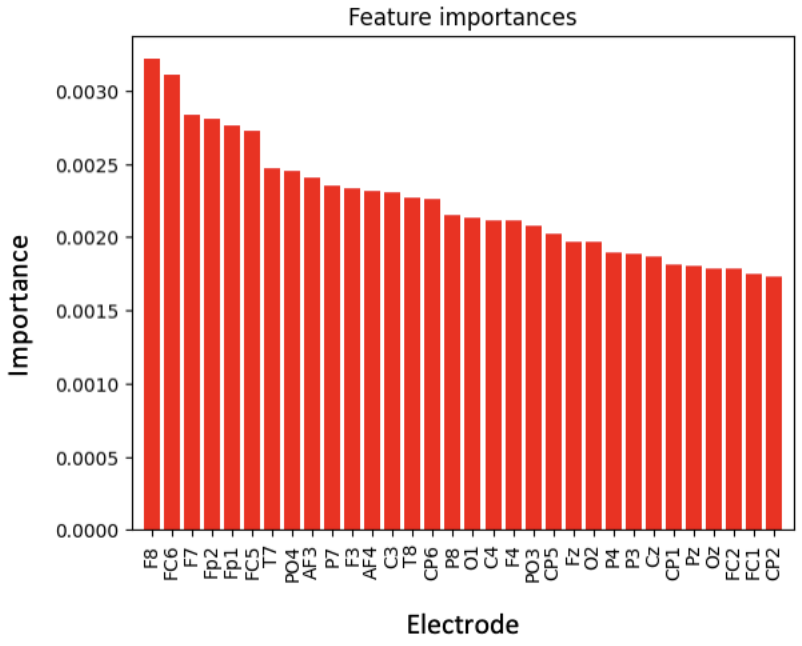


Figure 7.14: Three class classification sliding window electrode importance's

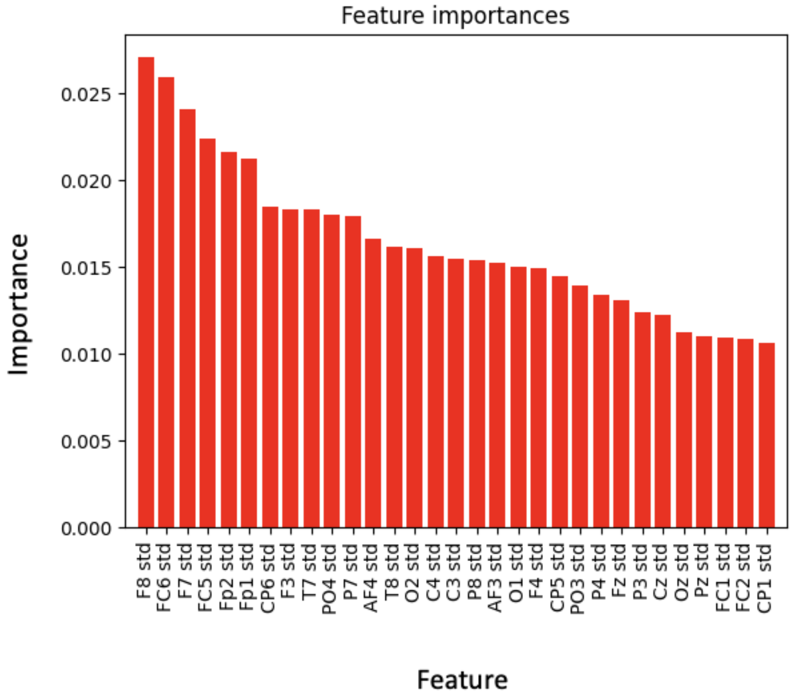


Figure 7.15: Three class classification sliding window feature importance's (>0.01)

7.3.2.5 ML Findings

Upon analysing our results, several recurring themes become apparent. One significant pattern involves the variance in accuracy observed when comparing the use of the entire sample against real-time classification. This distinction holds for both binary and three-class classifications. This is expected; our hypothesis behind this is two-fold. Firstly, the inability to ensure that the participants were at the correct level of engagement for every 5-second window of data could play a role in this disparity. Secondly, the limited amount of data the model can learn from in each sample significantly impacts its ability to make accurate classifications. These findings highlight the challenges imposed by temporal constraints and dataset size. This also highlights the complexities inherent in real-time classification scenarios.

Our pilot trials showed that time domain feature set outperformed both the frequency domain feature set and the combined time and frequency domain feature set. These performance improvements were more pronounced when a larger window of data was used (10 minutes). Discussion on why this may be the case can be found in section 7.3.2.6

Some insights emerge when comparing the complete (time domain) feature set with the selected feature set. Notably, in three out of the four tasks, the selected feature set outperforms its complete counterpart. Examining the consistently chosen features reveals clear patterns, which we look into in more detail in section 7.3.2.6. No discernible trends were indicating whether the selected feature set specifically enhanced precision or recall.

There is a disparity in model performance between the binary and three-class engagement classification tasks highlighting several potential challenges inherent in the multi-class scenario. First, the EEG features used may lack sufficient discriminative power to clearly distinguish between high, moderate, and low engagement levels, particularly if the physiological signals for moderate engagement overlap with those of high or low engagement. The responses to our questionnaire revealed substantial overlaps in participants' engagement scores during the moderately engaging segment of the experiment. This overlap is compounded by the inherent subjectivity and potential ambiguity in defining engagement levels, this may have led to imprecise labelling. The moderate performance observed in the three-class task, only slightly above random guessing, suggests that the features or model may capture only broad differences (e.g., high versus low engagement) but struggle with finer-grained distinctions.

Numerous papers exist in the realm of engagement recognition, although direct comparisons prove challenging due to inherent dissimilarities in stimuli, testing procedures, and data volume. As we have shown in our previous analysis, certain validation and data splitting methodologies, for example having data from a single participant in both train and test sets leading to data leakage often results in overestimation of model performance. Despite these variations, insights can be made. While focusing solely on one modality, our study still yields comparable results.

In a study by Chanel et al. (2008), the researchers focused on classifying three distinct levels of engagement using a similar procedure to what we used. The different engagement levels were elicited using Tetris's various difficulties as stimuli. Various physiological signals were recorded and used to develop their ML models however, EEG was not. An accuracy of 53.33% was achieved using an SVM. Similar to our results, the middle level of engagement was particularly hard to classify. The paper's validation methodology is unclear. While it states that participant data was excluded from both the validation and testing sets, there is no mention of cross-validation or reporting of standard deviation. This raises questions about whether the reported results were derived from a single participant, potentially limiting their generalisability to other participants. Alternatively, if cross-validation was used, the paper does not provide information about the typical range of accuracies observed.

Chanel et al. (2011) conducted a follow-up study that utilised EEG signals and the physiological signals recorded before. The addition of EEG data improved the model accuracy to 65%, which is comparable to the results we achieved using solely EEG data. This highlights the value of EEG signals when trying to classify levels of engagement. This also highlights that the models we have developed could be improved if more modalities are introduced. The paper mentions that nested leave-one-participant-out cross-validation was used; however, the standard deviation of the results is only presented visually in the charts and is not explicitly stated in the text.

Furthermore, in the work by Abadi et al. (2013), a combination of GSR, facial motion analysis, and Frontal EEG data was employed to classify high and low engagement during film viewing. Their approach achieved a 70% accuracy rate when classification was based on the intended emotional response to stimuli. However, when individual subject questionnaires guided the classification, accuracy improved to 76%. This approach suggests considering subjective participant responses could improve classification results, especially when participants' perceived engagement levels vary. Integrating these insights could potentially lead to improved results in our own models. This paper employs both cross-validation and leave-one-participant-out methodologies. The highest reported accuracy, 76% for a binary classification task using multiple modalities, was achieved by splitting the data into smaller segments and applying a majority vote. In comparison, the most similar procedure to ours used only EEG signals, applied the leave-one-participant-out methodology with the complete sample, and achieved an accuracy of 65%, compared to our 75% accuracy.

In a study by Sümer et al. (2023), engagement levels were tracked during student learning sessions. The study revealed that personally reported classifiers outperformed other methods, similar to what was found by Abadi et al. (2013). When tested on two sets of students, the best-performing set achieved an AUC of 0.72 using LSTM networks, with both LSTM and RF emerging as the top-performing classifiers. RF demonstrated classification accuracies for low engagement at 0.185, moderate engagement at 0.768, and high engagement at 0.608. In comparison, our best results achieved accuracy scores of 0.71, 0.29, and 0.57; there is a clear contrast in the level of engagement that both studies struggle to classify. This could be due to model development, feature extraction, etc., or it could more likely be due to the differences in stimuli and the engagement levels they evoke. This paper utilised a nested leave-one-participant-out validation approach.

In a separate study conducted by Apicella et al. (2022), a novel wearable system was developed for EEG-based engagement detection. This system assessed both cognitive and emotional engagement. Cognitive engagement aligns more closely with the engagement evoked in our study. Their study aimed to distinguish between high and low-engagement states. The most comparable "within subject" result achieved an accuracy of 0.77, which matched the accuracy obtained in our study. These findings underscore the potential of personal or integrated EEG systems. This research also conducted a cross-subject analysis, which is more comparable to our approach, achieving an accuracy of 73% compared to our 75%. For hyperparameter tuning, they stated that cross-validation was performed by leaving two trials out but each participant undertook eight trials, it is unclear whether this was applied to the entire dataset or just the training set. Additionally, they mention that the reported results are averaged across all participants, suggesting that a leave-one-participant-out methodology may have been used.

In summary, our research reveals that our ML model effectively detects engagement by utilising VR as a stimulus and relying solely on EEG signals. Notably, even with relatively limited data, our results demonstrate accuracy levels comparable to those reported throughout the literature, where diverse amounts of data and multiple modalities were employed.

Table 7.9: Engagement Recognition Literature

Authors	Title	Modalities Used	Model developed
Chanel et al. (2008)	Boredom, Engagement and Anxiety as Indicators for Adaptation to Difficulty in Games	GSR, Blood Pressure, Heart rate, Respiration and Temperature	3 Class Classification (53.33%)
Chanel et al. (2011)	Emotion Assessment From Physiological Signals for Adaptation of Game Difficulty	EEG, Peripheral Signals	3 Class Classification (63%)
Abadi et al. (2013)	Multimodal Engagement Classification for Affective Cinema	EEG, GSR, Facial Tracking	Binary Classification (71%)
Sümer et al. (2023)	Multimodal engagement analysis from facial videos in the classroom	facial videos	0.72 (AUC for three class classifications)
Apicella et al. (2022)	EEG-based measurement system for monitoring student engagement in learning 4.0	EEG	0.73 (Binary classification)

7.3.2.6 Feature Importance and Correlation Findings

Our analysis of feature importance identifies clear insights: the distribution and variation within the data emerged as pivotal factors in our classifications. Specifically, kurtosis, skewness, and standard deviation were identified as the most influential features. Notably, F8 and F7 consistently stood out as top features across the classification tasks, highlighting their substantial impact on our results. These electrodes lie over the inferior frontal cortex (Hoque & DelRosso, 2014). The inferior frontal cortex has key roles, including executive function (mental skills that include working memory, flexible thinking, and self-control) and social cognition (Diveica et al., 2023). It is clear to see how these functions may relate to engagement. These feature importance scores emphasise the importance of these brain regions in the context of engagement detection. Their consistent prominence underlines their important role in classifying varying engagement states, providing insights into the neural mechanisms behind human engagement responses.

Corelation analysis of the features and engagement level was performed. We used Pearson correlation to assess the strength and direction of the linear relationship between the EEG data and engagement level. Initially raw signals were used but there were no channels that had no strong correlations. Correlations for specific channels and features were then calculated.

Results for the statistically significant features can be found in table 7.10. The results of the correlation analysis between EEG features and engagement levels reveal some patterns. A moderate positive correlation was observed between the kurtosis of the EEG signal from the C3 channel and engagement, indicating that higher kurtosis values, which reflect more peaked or outlier-prone distributions, are associated with higher engagement. Similarly, the variance in the PO4 and CP1 channels exhibited moderate positive correlations with engagement levels, suggesting that increased signal variability in these regions corresponds to heightened engagement. In contrast, the mean signal values in the FC6 and C4 channels showed moderate negative correlations with engagement, indicating that lower mean amplitudes in these regions are associated with higher engagement.

Some comparisons can be made between this correlation analysis and the feature importance results. When looking at the binary and three class classification using the complete sample, C3 kurtosis is highlighted as a prominent feature in both the feature importance and correlation analysis. None of the other features highlighted by correlation analysis appeared in the feature importance. This could suggest that there is a non-linear relationship between the selected features and engagement.

Table 7.10: Feature Correlations

Feature	Correlation	P value
C3 Kurtosis	0.35	0.02
P04 Variance	0.31	0.04
CP1 Variance	0.31	0.04
FC6 Mean	-0.36	0.02
C4 Mean	-0.36	0.02

A difference becomes evident when comparing our approach to feature extraction and analysis with that of existing literature. While most previous studies predominantly rely on frequency domain features, mainly focusing on frequency bands, our methodology takes a different path. We concentrate on time-domain statistical features instead of the frequency domain because time domain features performed better in our pilot trials. This could be because the time domain features were able to observe the entire waveform rather than just individual frequency bands (Li et al., 2020) and therefore may have gained more information. More targeted frequency domain features (for example targeting certain frequency bands we know are relevant) may lead to improved performance. While more in-depth feature engineering, such as exploring complex frequency domain features, remains a potential avenue for enhancing our methodology, our study serves as a demonstration of the feasibility of employing computationally simple statistical features. The results obtained through this approach indicate its viability, showcasing that even with straightforward features, engagement detection is achievable.

7.4 Discussion

7.4.1 Research Objective I: Understand whether VR is a Reliable Stimulus for Engagement Recognition.

In section 6.3.1 (validation of dataset), we confirmed that VR is a reliable stimulus for engagement recognition. A one-way ANOVA test was performed on the results obtained from the GEQ to verify that the correct levels of engagement were evoked. The one-way ANOVA test confirmed a significant difference between the stimuli and their corresponding engagement ratings ($p<0.05$).

7.4.2 Research Objective II: Investigate if Engagement can be Detected within Virtual Environments using ML and Compare the Results to Literature and Studies that Utilise Non-immersive Stimuli.

The results achieved from our ML analysis were compared to literature that utilised non-immersive stimuli. While direct comparisons are hard to make, we found that our ML model effectively detects engagement by utilising VR as a stimulus and relying solely on EEG signals. Notably, even with relatively limited data, our results demonstrate accuracy levels comparable to those reported throughout the literature (see 6.3.2 for more detail).

7.4.3 Research Objective III: Investigate the Specific Features of Data that Indicate Engagement and Compare these to Affective Computing and Medical Literature

We compared our feature importance results to affective computing and medical literature, and overall, we found some interesting insights. We found similarities when comparing it to medical literature; for example, F8 (the most significant feature in our analysis) relates to social inhibition and emotional control. When looking at affective computing literature, most studies utilise frequency domain features, whereas our analysis focuses on time domain statistical features. The results obtained through this approach indicate its viability (see 6.3.2.6 for more discussion).

Chapter 8: Discussion and Conclusion

8.1 Research Questions Addressed

1. ***Can daily mental health throughout the course of therapy be monitored using off-the-shelf wearable sensors and ML techniques (Chapter 4)?***

We explored this research question in Chapter 4. In particular, this study examined whether the PHQ-9 levels of patients could be predicted daily using machine learning (ML) throughout the course of repetitive transcranial magnetic stimulation (rTMS) treatment. Fitbits were used to collect data. We were able to classify:

- When a patient would have reported being severely depressed (PHQ-9 ≥ 20) with 72% accuracy
- If a patient's PHQ-9 scores are higher, lower or the same as the previous day with an accuracy of 65%
- If a patient's PHQ-9 scores are higher, lower or the same as the baseline scores reported at the start of the course of therapy with an accuracy of 62%

Prior to this, no study had investigated whether ML can be used to track the depression levels of Treatment Resistant Depression (TRD) patients throughout the course of therapy; our results suggest that models can be developed to track populations of people that would be considered outliers among the healthy population. Because no prior study had investigated mental health monitoring in this population, we couldn't compare our results to the literature directly. Our results do compare comparatively well to what is reported in the literature, results in the literature that generally looked at a healthy population and predicted depression in less granularity (predominantly weekly and biweekly). Our results performed equally or improved on those results.

With this being the case, our results suggest that mental health can be monitored daily throughout therapy using off-the-shelf wearable sensors and ML techniques. Off-the-shelf sensors are not the gold standard; however, their accuracy is acceptable (Chevance et al., 2022; Haghayegh et al., 2019) and the devices are low cost. The accessibility of off-the-shelf wearable sensors allows for easier adoption in a clinical setting. Our results and findings serve as proof of concept for further development. Further research needs to be conducted to understand how these models could be implemented in a clinical setting. Our research looked into predicting PHQ-9 scores. ROC analysis revealed that the PHQ-9 exhibited an area under the curve of 0.95 when diagnosing major depression (Kroenke et al., 2001). Consequently, our predictions should extend beyond questionnaire scores to accurately reflect true depression status. We also need to consider what the ML performance benchmarks are for a system to be useful for clinicians and patients.

2. ***Is VR a reliable stimulus for emotion (Chapter 5) and engagement (chapter 7) elicitation throughout virtual therapy?***

This research question was investigated in two separate studies; in the first, we looked to understand if VR was a reliable stimulus to elicit different emotions and in the second, we explored whether VR was a good stimulus to elicit different levels of engagement.

In Chapter 5 data was collected while users were experiencing various virtual environments selected to elicit one of four different emotions corresponding to the four CMA quadrants. To validate this data, statistical and ML analysis was performed. Statistical analysis confirmed that users felt four distinct emotions during the intended virtual experiences. ML analysis was compared to other well-known affective datasets that utilised non-immersive stimuli. The baseline ML results reported in chapter 5 outperformed all other baseline results reported in the other established datasets. The ML results are explored in more depth in research question 3. With the statistical and ML analysis, our results suggest that VR is a reliable stimulus for emotion and could be superior to non-immersive stimuli.

As for engagement, a similar analysis was carried out in Chapter 7, both statistical and ML. The statistical analysis revealed there was a significant difference between the different stimuli and the participants corresponding engagement ratings ($p<0.05$). This suggests that the stimuli evoked the intended engagement level and that the three different engagement levels were clearly distinguishable. The ML results were once again hard to compare to literature as there are no direct comparisons; however, our results are comparative to other similar studies. Taking both the statistical and ML analysis into account, we have shown VR to be a reliable stimulus for engagement elicitation. We believe VR has shown to be a good stimulus for emotion and engagement and to build corresponding ML models because the user is able to engage with affective stimuli more profoundly than non-immersive stimuli due to the increased feeling of immersion and presence (Riva et al., 2007; Baños et al., 2004). Correlations in physiological dynamics between real-world and immersive environments have also been observed, and VR has been shown to be the most similar to the real-world terms elicitation of physiological response (compared to photos and 360 panoramas) (Higuera-Trujillo et al., 2017). This further highlights the potential of immersive environments to evoke emotions similar to what would be evoked in the real world.

3. ***Can accurate ML models be developed to recognise engagement and emotions during immersive virtual experiences (Chapter 5 and 7)?***

Chapter 5 investigated and tested an engagement recognition model, and Chapter 7 investigated and tested an engagement detection model, both of which were developed using data collected from immersive virtual experiences.

We investigated three different classification tasks regarding emotion detection:

- Four class classification of the four CMA quadrants (high arousal/high valence, high arousal/low valence, low arousal/high valence and low arousal/low valence
- Binary classification of high and low arousal
- Binary classification of high and low valence

The results obtained from these tasks were compared to those of other related studies, all of which utilised non-immersive stimuli. Our results outperformed all other baseline results reported in established affective datasets. This highlights that robust ML models can be developed to recognise emotion and that immersive stimuli could be better than non-immersive models for developing these models. We believe this to be the case because of the increased ability of immersive virtual environments to elicit emotion and physiological responses resulting in higher quality data. These initial baseline models were developed using established methods partially as a way to compare to other studies that have used similar methods. This means that the main difference between the studies was the data and the stimuli used to collect it (immersive vs non-immersive).

We further investigated emotion recognition models in chapter 6 where we used deep learning techniques. This research focused on the four-class classification task. This chapter demonstrated that deep learning models can outperform shallow models for emotion recognition within immersive virtual environments. The deep learning model achieved an accuracy, precision, recall and f1 of 0.743, 0.743, 0.750 and 0.743, respectively, compared to the shallow methods, 0.719, 0.76, 0.72 and 0.72. The deep learning model used raw physiological data, whereas the shallow methods relied on extracted features. This could be a reason for the improved performance; the deep learning model may have been able to gain information from segments of the signal that were not captured in the extracted features.

When investigating ML models that detect various engagement levels, we looked at 4 different classification tasks:

- Binary classification of the complete sample: classification of high and low engagement using the complete sample to extract features
- Three class classifications of the complete sample: classification of high, medium and low engagement using the complete sample to extract features
- Binary classification of sliding windows: classification of high and low engagement using a sliding window with a size of 5 seconds and a step of 1 second
- Three class classifications of sliding windows: classification of high, medium and low engagement using a sliding window with a size of 5 seconds and a step of 1 second

These classification tasks can provide information on how engaged in an immersive environment someone is. Binary classification gives a simple measure of whether or not one is engaged. Three class classifications can additionally provide information on whether a user is moderately engaged; however, as expected, this comes at the expense of accuracy. There is limited literature to compare to in this field so presenting both binary and three class classification allows a more thorough comparison to literature increasing the validity of the models presented. This information in a clinical context could provide more enhanced and personal care by allowing treatments to be tailored to increase each user's engagement. With engagement playing a role in effective therapy (see 2.4 for further discussion), the ability to understand patient engagement and, in turn, create more engaging therapy sessions could improve patient experience and outcomes.

Our results suggest robust ML models can be developed to classify levels of engagement during virtual immersive experiences. Our results compare well to similar studies, especially considering that the other studies utilised multiple modalities, whereas ours relied on a single modality (EEG). These findings were consistent across all classification tasks: binary, three-class, complete sample, and sliding window. Using purely EEG data allows for engagement detection within combined VR and EEG headsets, which are starting to be developed (Ag, 2023). This would result in a simple-to-use all-in-one device, which would be essential for acceptability and use within real-world clinical settings where ease of use and limiting the number of separate devices needed are important.

4. ***Can real-time emotions of VR users be detected, and how can Explainable Artificial Intelligence (XAI) provide interpretable explanations for these predictions? (Chapter 5)?***

This research question was answered in Chapter 6 and was an expansion of the results presented in Chapter 5. In order to create a model to detect real-time emotion, we utilised the data from Chapter 5 and carried out a sliding window to segment the data into 5-second chunks of data. A 5-second window is commonly used throughout literature, and we also tested 2.5 and 10-second window sizes, both of which a 5-second window outperformed. As this created a larger dataset with drastically more instances, we could use deep learning. The results obtained using a deep learning algorithm outperformed the results presented in Chapter 5. There are multiple benefits of using deep learning techniques: typically an improvement in accuracy, its ability to process raw data, negating the need for time-consuming manual feature extraction (Janiesch et al., 2021) and the possibility of knowledge gain. One way in which knowledge can be gained is by allowing the model to find its own features, some of which may not have been considered in traditional feature extraction. We found that our model was gaining information from the ST segment of the ECG wave, which is not considered a common feature to extract. There are, however, also some negatives to a deep learning approach: these models can be time-consuming to train and computationally expensive to run when compared to shallow methods. They can also be considered a black box; however, in our research, we have provided a methodology to begin to uncover the black box. Our results show promise that VR users' real-time emotions can be detected using raw time series physiological signals and deep learning techniques.

For the XAI section of this research question, we employed an implementation of SHAP to provide local explanations in the form of visualisations of the importance of each segment of the raw time series physiological signals. These visualisations allowed us to interrogate the model outputs and conclude that the model was utilising established features of ECG, GSR and eye tracking data. For the ECG signals, we found that all participants shared common heart rate themes across the videos. Like affective computing and medical literature, the QRS complex, specifically the R peak, was an important feature of the deep learning model. The ST segment also a relatively important feature despite not commonly being highlighted in literature. We found that GSR was consistently the least impactful modality we analysed for the deep learning model, this is in contrast to the feature importance reported from shallow methods. We deduced that this was probably because key GSR features couldn't be captured with a 5-second segment of data. We found that small, quick movements were important to the deep learning model. Fixations also seemed to be an important feature for the deep learning model. Both of the findings are consistent with the literature and the results reported in Chapter 5. Further insights on all three modalities can be found in section 6.3.2.2. These findings improved the confidence that the model had learnt valuable patterns. In the global explanations, we investigated the overall importance of each modality. This uncovered that the impact of modality is more dependent on the specific emotion rather than the individual participant. It was also found that the importance of each modality significantly differed between the deep learning model and the shallow methods; ECG played a much larger role within the deep learning model, and GSR played a significantly lower role. Further global analysis can be found in section 6.3.2.1.

5. ***Can the physiological signals and their specific features that are indicators of emotion and engagement be identified? (Chapters 4, 5, 6 and 7)?***

This research question was answered throughout all project chapters of this thesis. We identified key themes and modalities that contributed to emotion and engagement recognition models.

For emotion recognition models, our feature importance analysis started in Chapter 5, where we analysed the feature importance's of the ECG, GSR and eye-tracking extracted features. The importance's of these were compared against affective computing and medical literature. In general, the feature importance's was consistently used and validated across both areas of literature. We were able to identify biological reasons behind the importance of certain physiological responses. This provides greater confidence and validates the models and results presented. These initial importances provided in Chapter 5 provided a baseline to compare the XAI results reported in Chapter 6. Multiple key insights presented in Chapter 5 validate the model that was presented by highlighting that the model had learnt representations of the data similar to what has been reported in affective computing and medical literature.

Overall, we found that the key indicators of emotion were generally agreed upon across shallow and deep learning models. These include the maximum amplitude of the ECG signal, various features that can be extracted from R peaks and micro-saccades and fixations of eye-tracking data. We also found that ECG data was more impactful on the deep learning model, suggesting there may be data we did not capture in our ECG feature extraction in Chapter 5. One of these possible features is the ST wave. The local explanations in Chapter 6 indicate that this portion of the QRS complex played an important role in the deep learning model, but no features related to the ST wave were extracted in Chapter 5. For GSR data, we identified that larger time periods (more than 5 seconds) of time series data are required to provide optimal information to the model. This was evident in the reduced role GSR data played in the model that used sliding windows. From the visualisations from the local explanations, we were able to confirm that key features reported in Chapter 5 (number of peaks and valleys, etc) were not captured in short periods of data. This resulted in an overall reduction in the importance of the GSR signal and less noticeable patterns of importance in the signal.

For engagement, our analysis is limited to one modality, EEG. We were able to identify electrodes that were key indicators of engagement. In particular, F8 and F7 consistently stood out as top features. We were able to relate these electrode placements to brain regions that play a role in engagement. These electrodes lie over the inferior frontal cortex (Hoque & DelRosso, 2014). The inferior frontal cortex has key roles, including executive function (mental skills that include working memory, flexible thinking, and self-control) and social cognition (Diveica et al., 2023). We were also able to highlight that time domain features such as the mean standard deviation and skew of each channel, while not commonly used throughout literature, could be valuable in classification tasks similar to what we presented.

Overall, we learnt that all modalities we used were useful in emotion and engagement detection. Our analysis of emotion detection uncovered that there doesn’t seem to be one best modality and that the importance of modalities can be model-dependent. However, It was clear that a combination of modalities improved model performance. We did not explore multiple modalities for engagement recognition, but we would hypothesise that introducing more modalities could improve model performance as this is a similar task.

8.2 Contributions

The research presented in this thesis indicates the effectiveness of immersive VR-based stimuli in eliciting emotions compared to non-immersive stimuli. This discovery not only enriches our comprehension of emotional responses but also underscores the potential of VR technology in psychotherapeutic contexts. Furthermore, it has broader implications:

- Affective Computing Community: This research underscores VR as a promising avenue for enhancing emotion recognition systems. Showcasing VR's ability to evoke more pronounced emotional responses due to the increased immersion and presence felt compared to non-immersive stimuli and create more pronounced data. Our research suggests integrating VR into affective computing frameworks could lead to more accurate and robust emotion recognition algorithms.
- Medical Community and Virtual Immersive Therapy: The demonstrated feasibility of effective emotional elicitation in VR holds implications for the medical community, particularly in the realm of virtual immersive therapy. By highlighting VR's capacity to evoke and modulate emotions in a controlled environment, this research suggests that virtual immersive experiences could enhance therapeutic interventions by evoking more realistic and impactful emotions in patients (Riva et al., 2007; Baños et al., 2004). This is especially pertinent in therapies like exposure therapy, where the deliberate evocation of specific emotions is crucial for therapeutic outcomes.

In this thesis, the feasibility of daily mental health monitoring of patients diagnosed with treatment resistant depression (TRD) throughout the course of therapy with off-the-shelf sensors was established. The practicality of continuous mental health monitoring throughout the course of therapy using readily available sensors was demonstrated. The research paves the way for more accessible and cost-effective solutions in mental health care. This contribution has implications for:

- Medical Community: This research emphasises the possibility of implementing low-cost remote monitoring for patients undergoing therapy. Currently, the monitoring of patients is limited in clinical practice despite the evidence it can improve patient outcomes (Jensen-Doss et al., 2016). This research highlights a potential avenue for enhancing patient care and outcomes through continuous tracking throughout the treatment process overcoming some of the current limitation in patient monitoring mentioned above.

This thesis carried out the identification of Machine Learning Models and Methods for Emotion and Engagement Recognition during immersive virtual experiences. This research identified and evaluated various machine learning models and techniques specifically tailored for emotion and engagement recognition. Analysis and recommendations regarding model development and feature engineering were made throughout. The thesis contributes to the advancement of emotion and engagement recognition technology. This contribution benefits the following community:

- Affective Computing: This research provides evidence supporting the detection of emotions and engagement levels within immersive virtual experiences. Throughout the thesis, we have emphasised insights into methods and features crucial for successful model development. Additionally, we present baseline results for the VREED dataset and demonstrate improved real-time outcomes through the implementation of deep learning techniques. As there are very limited affective datasets containing data collected using immersive stimuli, these results were one of the first to compare emotion recognition using immersive vs non-immersive stimuli. The results highlighted the possible benefits of utilising immersive stimuli for emotion and engagement recognition; this could further improve affective models in the future.

This thesis demonstrated a real-time emotion detection model and XAI framework that utilised raw time series physiological signal data. The thesis introduces a deep learning-based real-time detection model for accurately identifying emotions. Additionally, it proposes an XAI framework that detects these emotional states and provides interpretable visual results. The XAI results were used to validate the model and provide insights into further model development and feature engineering of ECG GSR and eye-tracking data. This contribution has implications for the following communities:

- Affective Computing: Our research demonstrates that deep learning methods can outperform shallow methods in emotion detection during immersive virtual experiences, offering insights for future model development. This could be due to the increased amount of data in each datapoint; it could possibly be using features that aren't captured in manual feature extraction. Shallow methods could still provide benefits when there is limited data or if computational efficiency is a concern. In addition, we also introduce a model and modality-agnostic XAI framework, enabling the explanation of emotional states detected from raw time series signals. This methodology has the potential to not only enhance the accuracy of emotion recognition systems but also foster transparency and interpretability, which are crucial for their acceptance and utilisation in practical settings.
- Medical Community: By showcasing the potential for real-time emotional feedback during therapy sessions, our work opens up new avenues for supporting clinicians in clinical decision support and personalised medicine. One example is in therapy; the underlying physiological data can support and inform the therapist and support further direction and decisions of the therapy. An example where a system like this could support decisions and

provide more personal care is in scenarios where patients may engage in social masking; in such situations, the clinician may not be able to identify the true emotion being felt and, therefore, potentially make misinformed decisions. Trust and support for AI systems could also be improved via XAI, because, the proposed XAI framework allows clinicians to comprehend the underlying reasons behind predictions made by black-box methods on physiological signals. This transparency has the potential to allow clinicians to make informed decisions and adapt therapeutic interventions more effectively, ultimately improving patient outcomes.

This research presented insights into physiological signals, their significance in emotion and engagement, and a comparison to medical literature. The thesis investigates physiological signals and their underlying features and mechanisms. By exploring the importance of specific physiological signals and their components in the context of emotion and engagement, the research provides guidance for feature engineering for affective model development. The interrogation of the importance of the underlying feature also gave context to how the models worked, providing further confidence that they are learning clinically sound representations of the physiological data. This contribution has implications for the following communities:

- Medical Community: Our research offers insights into the requirements of immersive virtual therapy systems, outlining components necessary for effective therapeutic interventions. We provided guidance for the development of immersive virtual therapy platforms that cater to the specific needs of patients and clinicians. Additionally, our work contributes to the development of models that are easily comprehensible to clinical staff, facilitating their integration into medical practice. Moreover, our efforts towards creating transparent and explainable models align with the broader goal of providing medical practitioners with tools that enhance decision-making processes and improve patient care outcomes.
- Affective Computing: Our research provides guidance for the further development of affective models and feature selection techniques within affective computing. By identifying key considerations and methodologies, we contribute to advancing the further development of emotion recognition systems. Our efforts in this area aim to pave the way for more effective and reliable applications of affective computing.

This research presented an analysis of ML training and validation techniques for emotion and engagement recognition. The thesis investigates a variety of commonly used methods, demonstrating some methods tendencies tendency to overestimate performance metrics. By advocating for robust evaluation frameworks, such as subject-independent or leave-one-subject-out validation, the research offers a pathway to more accurate, easily comparable and generalisable assessments of model performance. These contributions have implications for the following communities:

- Affective computing: Our results demonstrate that using a holdout set can lead to both overestimation and underestimation of performance metrics as seen through chapters 4 and 5, depending on the specific composition of the holdout data. Furthermore, employing standard cross-validation techniques—where data from a single participant can appear in both the training and validation sets can result in overestimated performance metrics. These effects are less pronounced when participant data is included in both training and testing sets, provided that individual trials remain separated. However, the research strongly suggests that nested leave-one-participant-out cross-validation is the most robust validation method. When this is not feasible, a holdout methodology involving multiple participants, with no overlap of participants or trials between the training and testing/unseen sets, should be employed. Finally, the thesis highlights the importance of exercising caution when comparing results to the literature, given the variability in validation techniques employed across studies.

The thesis has advanced the aim of enabling at-home virtual therapy through several contributions:

- Emotion and Engagement Elicitation in VR: By demonstrating the efficacy of eliciting emotions and engagement in VR, the thesis establishes a foundation for immersive virtual therapy. This is essential in therapy, where engaging patients and evoking emotions play a key role in patient outcomes (Gaston, 1998; Gomes-Schwartz, 1978; Zuroff et al., 2016; Öst et al., 1997).
- Real-Time Tracking of Emotion and Engagement: The thesis showcases the ability to track emotion and engagement levels in real-time. This real-time feedback could provide clinicians with valuable contextual information, allowing them to tailor interventions more effectively to the patient’s needs and overcome some existing limitations, such as social masking.
- Continuous Monitoring of Mental Health States: The thesis utilises off-the-shelf hardware to enable cost-effective and continuous monitoring of mental health states throughout therapy. This capability extends therapy beyond the confines of clinical settings, allowing continuous patient mental health monitoring in their daily lives.
- Insights into Model Features: By providing insights into the features used to develop emotion and engagement recognition models, the thesis enhances the trust and

transparency of these models. Understanding the underlying features improves the credibility of the models and informs future model development.

- XAI Methodology: The identification of a modality and model-agnostic XAI methodology represents an advancement. This methodology offers explanations for time series data in a visually interpretable manner, enhancing the transparency and interpretability of the models. Clinicians and patients can gain insights into why specific predictions are made, fostering trust in the technology.

Collectively, these contributions address various aspects of at-home virtual therapy, from emotion and engagement elicitation to continuous monitoring and explainable model development. By providing tools and insights across these domains, the thesis lays the groundwork for the future research of effective and accessible immersive virtual therapy solutions. Below we explore how the systems proposed in this thesis could be used in real world clinical systems.

The remote monitoring system described in Chapter 4 can be used as a clinical decision support tool, enhancing therapy and treatment adjustments based on depression classifications and clinical observations. By providing continuous data through an online dashboard, this system could supplement clinician observations with additional context, especially on days without direct patient contact. Identifying outlier data could facilitate more timely consultations.

For emotion and engagement detection, during in-person therapy sessions, these systems can offer decision support by providing contextual data on patient emotions, aiding clinicians in adapting therapy when social masking occurs. Visual explainable results can enhance confidence in the system's predictions. Similarly, engagement recognition can help clinicians tailor sessions to maintain patient interest.

In remote therapy sessions without a clinician present, these systems can support immersive virtual therapy. These sessions could be supplementary and additional to existing treatment programs. Sessions can be personalised based on emotional and engagement responses, with all data fed back to clinicians for review and subsequent therapy adjustments.

Overall, these systems contribute to a more comprehensive treatment approach, enabling continuous monitoring and adaptation beyond traditional in-person sessions.

One example of where this system could improve treatment outcomes is in the treatment of PTSD. The variability of treatment timing can vary significantly due to sessions being cancelled or rescheduled and it has been found that frequent scheduling of sessions can maximise the treatment outcomes (Gutner et al., 2016). If at home therapy was available, cancelled sessions could be undertaken whenever suits the patient irrespective of the clinician. Research has shown that standalone self-guided or automated therapy (no clinician present) is effective (Gaerets et al., 2021). Cognitive behavioural therapy and exposure therapy are both common treatments for PTSD. Both of these can be enhanced and more personalised when the patients emotion throughout is considered (Samoilov & Goldfried, 2000). Understanding engagement throughout this automated/self-guided therapy can also lead to more personalised, engaging and effective treatments. On top of replacing missed therapies, more frequent supplementary sessions could also be prescribed with this technology. The daily monitoring of mental health may also allow the prioritisation of patients and who need treatment the most.

8.3 Limitations

In this thesis, several limitations warrant consideration in interpreting the findings and implications of the research. Firstly, the scale of the studies conducted were relatively small datasets in comparison to larger-scale repositories that include extensive participant numbers and stimuli varieties. While expanding the scale of data collection could bolster the robustness of our conclusions, achieving this would entail substantial investments of effort, time, and resources. Nonetheless, within the confines of our dataset sizes, we were able to effectively demonstrate the efficacy of VR for emotion and engagement recognition, culminating in the development of robust ML models capable recognising various emotions and engagement levels within immersive environments.

Moreover, the participant population predominantly consisted of healthy individuals across Chapters 5, 6, and 7. While this allowed for a foundational understanding of emotional responses and engagement within VR settings, it inherently limits the generalisability of our findings to broader populations, including those with varying mental health conditions. Despite recognising the importance of diversifying participant demographics, logistical constraints such as limited time, access, and resources restricted our ability to collect data from more diverse cohorts. However, it's noteworthy that Chapter 3 did incorporate a study involving patients diagnosed with TRD, yielding encouraging results that warrant further exploration and replication in larger samples.

Furthermore, the absence of consideration for social demographic and cultural differences among participants poses another limitation. Our studies were conducted without accounting for the potential influence of varying social backgrounds and cultural contexts on emotional responses and engagement levels within VR environments. For example, different cultural norms may influence how individuals from diverse backgrounds perceive and engage with VR content, potentially affecting their physiological responses. As such, the generalisability of our findings to populations with diverse socio-cultural backgrounds may be limited, highlighting the need for future research to address this gap.

Additionally, the lack of a clinical setting in Chapters 5, 6, and 7 poses a limitation in extrapolating the applicability of our findings to real-world therapeutic contexts. While our research lays foundational groundwork for understanding emotional responses and engagement within immersive virtual environments, further refinement and validation within clinical settings are imperative to ascertain the practical utility of our findings in therapeutic interventions.

Finally, the generalisability of our findings regarding emotion engagement within immersive virtual experiences may be constrained by the specific stimuli utilised, which predominantly comprised of virtual experiences/videos and gaming contexts. Further investigation into the transferability of our findings across a broader spectrum of non-game and video based stimuli is warranted to ascertain the robustness and applicability of our conclusions in diverse immersive virtual environments.

8.4 Future Work

When looking ahead from our research that demonstrated the efficacy of mental health tracking throughout daily life, there are many avenues of research that could be explored, one of which being the expansion of a similar system into clinical contexts. The results showing that Fitbit data analysed using machine learning techniques can predict depression severity indicates the potential value of wearable activity trackers in monitoring depression and providing feedback on depression change. Fitbit data analysis could be combined with other data sources such as smartphone data analysis to improve the models developed here. Analysis that includes smart phone data has been found to reliably classify five-level human emotions with up to 95% accuracy (Kanjo et al., 2019). In the future, people with experience of depression could have ownership of an effective, unobtrusive depression monitoring and feedback system, widely available at low cost, to self-manage symptoms, find out what affects mood, take actions to improve mood, and seek help when needed. With dashboard visualisation, clinicians could observe depression changes over time, explore what works to improve symptoms and more effectively engage with, advise, and treat patients, allowing effective assessment of the impact of prescribed treatment. The system could also measure the impact of new interventions in clinical research; further system development is required. However, our findings are preliminary and require further research in much larger numbers of patients; research on feasibility and acceptability of use is also needed. Physical activity and sleep are just two aspects of a complex range of symptoms associated with the experience of depression.

An area to expand upon regarding the validation of immersive virtual experiences in emotion elicitation and emotional ML models would be to utilise EEG data, which many established datasets have used and shown to be a good predictor of emotion (Katsigiannis & Ramzan, 2018; Koelstra et al., 2012; Soleymani et al., 2012). EEG signals were not included in the VREED dataset due to the difficulties in collecting such data whilst wearing the specific model of VR head-mounted display that was used in this study without a significant redesign of the hardware. To overcome this, a reliable and compact EEG sensor that can be used in conjunction with a VR head-mounted display needs to be identified and/or developed. For instance, there have been significant development efforts in designing flexible sensors that capture physiological signals, including EEG for brain-machine interactions (Mahmood et al., 2019), and there has been pioneering work integrating such sensors with VR headsets (Mishra et al., 2020). In addition, some of the newer VR head-mounted displays are less bulky, and we were able to use them with EEG sensors later in chapter 7.

The data and models produced in this thesis could be further used within real world applications. The models presented could lead to more effective communication and improve applications that rely on human-computer interaction in many domains (Hassouneh et al., 2020; Lane & D'Mello, 2018; Mohd et al., 2019), including a variety of healthcare applications. For instance, an important use of affective systems is to assist parents, teachers and carers of children with autism (Picard, 2009). A reliable affective model, in which VREED could aid in the development of, could be used to detect the children's emotional states when they might not be otherwise visible and communicate this to the children themselves or others. This type of system could enhance communication and help assist social interaction. Affective systems have also been used as diagnostic and treatment tools in various stress and anxiety disorders. Specifically, PTSD is one of these disorders in which the diagnostics and treatment have benefited greatly from affective systems where virtual environments have been used to elicit stress in a controlled environment (Reger et al., 2011; Rizzo et al., 2009). The ability to detect and record affective states within these virtual environments may improve treatments by allowing clinicians to personalise the environment to the level that helps the patient feel the correct emotion or intensity of emotion (Yannakakis., 2018). The future works rely on data being collected from different populations and an increased scale of data collection.

There are many future advancements in this research concerning our investigations into deep learning techniques, XAI, and its application in real-world systems. To advance the model, the size of data and generalisability to a larger population need to be addressed. Two approaches could be taken to address this: collecting more data and using transfer learning. Transfer learning has already successfully been used for emotion recognition (Ng et al., 2015) and could improve model performance and generalisability. There has been much interest surrounding transformers, a type of deep learning architecture often used for tasks such as natural language processing and image generation (Lin et al., 2022). We focused on CNN-LSTM hybrid architecture as there is vast support in the literature for hybrid architecture and pure CNN and LSTM architectures, respectively. A CNN, in particular, is also more suited to extracting spatial features from sequential data than a transformer, which can be useful for capturing local patterns such as the waveform in an ECG. However, now that we have established the viability of deep learning models on this data, transformers could possibly be an avenue for enhancing performance.

For the XAI results, we looked at random segments of signals, It may be beneficial in the future to examine examples where a signal had a relatively high or low Shap value and then analyse the signal and its components. This could further identify which components make more impactful contributions to predictions.

Concerning our research into engagement recognition within immersive virtual experiences (chapter 7), several key strategies have been identified to enhance the ML results in our study. Our research focused on making a clinically viable system; however, if we wanted to pursue purely improving accuracy, expanding beyond EEG signals alone would help; by incorporating additional modalities, improved performance could be expected and has been shown in many other studies. Additionally, further exploration in the feature engineering process and exploring more into frequency domain features, presents an avenue for improving performance, however, this would come at a computational cost which may detract from the clinical viability. Furthermore, the application of deep learning techniques, exemplified in chapter 5 of this thesis, showcases promising advancements in various domains. Leveraging these advancements in the context of our research could further improve the effectiveness of our ML approaches.

The potential directions for this research are promising. Integrated sensors offer an avenue for exploration. These sensors, designed to integrate with VR headsets, pave the way for all-in-one systems that are easy to set up. Such advancements hold the promise of enabling immersive virtual therapy to be conducted in the comfort of one's home. Moreover, the concept of reactive experiences is also a topic of further research. By tailoring virtual experiences and therapies based on individual user responses and emotions, therapy can become highly personalised and effective. This approach not only enhances the user experience but also facilitates therapy without the need for a clinician's physical presence, allowing increased accessibility and personalised mental health support.

8.5 Conclusion

In conclusion, mental health stands as a significant societal challenge, with accessibility to therapy a concern. Barriers such as clinician shortages, geographical constraints, and patient reluctance to leave their safe spaces for therapy sessions underscore the need for innovative solutions. At-home therapy presents a promising avenue for overcoming these barriers, offering convenience and comfort to patients.

Our research delves into the potential of at-home therapy, particularly leveraging VR and immersive technologies to enhance therapeutic experiences. Through our investigations, we have explored several facets. Firstly, we have demonstrated the feasibility of continuous monitoring to track progress and mental health states throughout the course of therapy. Additionally, by demonstrating the effectiveness of VR in eliciting appropriate emotional responses and engagement levels, we provide a foundation for the integration of immersive therapy approaches. Furthermore, we provide analysis and insights into the underlying feature importance of our machine learning models, enhancing their transparency and improving trust among clinicians and patients. Lastly, our implementation of XAI enables clinicians to understand black box models analysing raw time series physiological data, facilitating informed decision-making.

Collectively, these advancements lay the groundwork for the development of immersive virtual therapy solutions. We hope this research can be used to refine further and expand various applications within healthcare, specifically at-home immersive virtual therapies.

Bibliography

"Evidence-based practice in psychology." (2006) *American Psychologist*, 61(4), pp. 271–285. Available at: https://doi.org/10.1037/0003-066x.61.4.271.

A plan for Digital Health and Social Care (2022) *GOV.UK*. Available at: https://www.gov.uk/government/publications/a-plan-for-digital-health-and-social-care/a-plan-for-digital-health-and-social-care (Accessed: 14 December 2023).

Abadi, M.K. *et al.* (2013) 'Multimodal Engagement Classification for affective cinema', *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* [Preprint]. doi:10.1109/acii.2013.74.

Abadi, M.K. *et al.* (2015) "DECAF: Meg-based multimodal database for decoding affective physiological responses," *IEEE Transactions on Affective Computing*, 6(3), pp. 209–222. Available at: https://doi.org/10.1109/taffc.2015.2392932.

Abdessalem, H.B. *et al.* (2019) "Toward real-time system adaptation using excitement detection from eye tracking," *Intelligent Tutoring Systems*, pp. 214–223. Available at: https://doi.org/10.1007/978-3-030-22244-4_26.

Abdul-Mageed, M. and Ungar, L., 2017. EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),*.

Acharya, U.R. *et al.* (2015) "A novel depression diagnosis index using non-linear features in EEG signals," *European Neurology*, 74(1-2), pp. 79–83. Available at: https://doi.org/10.1159/000438457.

ACQKNOWLEDGE data acquisition and Analysis Software - win: ACK100W, ACK100M: Research: BIOPAC (2020) *BIOPAC Systems, Inc.* Available at: https://www.biopac.com/product/acqknowledge-software/ (Accessed: May 4, 2020).

Adamakis, M. (2017) 'Comparing the validity of a GPS monitor and a smartphone application to measure physical activity', *Journal of Mobile Technology in Medicine*, 6(2), pp. 28–38. doi:10.7309/jmtm.6.2.4.

Ag, N. (2023) *NEUROSPEC AG - Research Neurosciences*, *NEUROSPEC AG Research Neurosciences*. Available at: https://www.neurospec.com/Products/Details/1077/dsi-vr300 (Accessed: 13 October 2023).

Agrafioti, F., Hatzinakos, D. and Anderson, A.K. (2012) “ECG pattern analysis for emotion detection,” *IEEE Transactions on Affective Computing*, 3(1), pp. 102–115. Available at: https://doi.org/10.1109/t-affc.2011.28.

Ahmad, Z. and Khan, N. (2022) ‘A survey on physiological signal-based emotion recognition’, *Bioengineering*, 9(11), p. 688. doi:10.3390/bioengineering9110688.

Aïm, F. *et al.* (2016) ‘Effectiveness of virtual reality training in orthopaedic surgery’, *Arthroscopy: The Journal of Arthroscopic &amp; Related Surgery*, 32(1), pp. 224–232. doi:10.1016/j.arthro.2015.07.023.

Albakri, G. *et al.* (2022) “Phobia exposure therapy using virtual and augmented reality: A systematic review,” *Applied Sciences*, 12(3), p. 1672. Available at: https://doi.org/10.3390/app12031672.

Alghowinem, S., Goecke, R., Wagner, M., Parker, G. and Breakspear, M., 2013. Eye movement analysis for depression detection. *2013 IEEE International Conference on Image Processing*,.

Alhargan, A., Cooke, N. and Binjammaz, T. (2017) ‘Multimodal affect recognition in an interactive gaming environment using eye tracking and speech signals’, *Proceedings of the 19th ACM International Conference on Multimodal Interaction* [Preprint]. doi:10.1145/3136755.3137016.

alive. 2020. *Emotions And Physiology | Alive*. [online] Available at: <http://www.alive.com/health/emotions-and-physiology/> [Accessed 25 June 2020].

Andreu-Perez, J., Leff, D., Ip, H. and Yang, G., 2015. From Wearable Sensors to Smart Implants-–Toward Pervasive and Personalized Healthcare. *IEEE Transactions on Biomedical Engineering*, 62(12), pp.2750-2762.

Angermueller, C. *et al.* (2017) ‘DeepCpG: Accurate prediction of single-cell DNA methylation states using Deep Learning’, *Genome Biology*, 18(1). doi:10.1186/s13059-017-1189-z.

Antoniadi, A.M. *et al.* (2021) 'Current challenges and future opportunities for XAI in machine learning-based Clinical Decision Support Systems: A systematic review', *Applied Sciences*, 11(11), p. 5088. doi:10.3390/app11115088.

Apa Dictionary of Psychology (2022) *American Psychological Association*. American Psychological Association. Available at: https://dictionary.apa.org/arousal (Accessed: August 24, 2022).

Apa Dictionary of Psychology (2022) *American Psychological Association*. American Psychological Association. Available at: https://dictionary.apa.org/valence (Accessed: August 24, 2022).

Apicella, A. *et al.* (2022) 'EEG-based measurement system for monitoring student engagement in learning 4.0', *Scientific Reports*, 12(1). doi:10.1038/s41598-022-09578-y.

ARGYLE, M.I.C.H.A.E.L. *et al.* (1970) "The communication of inferior and superior attitudes by verbal and non-verbal signals*," *British Journal of Social and Clinical Psychology*, 9(3), pp. 222–231. Available at: https://doi.org/10.1111/j.2044-8260.1970.tb00668.x.

Ashley, E.A. and Niebauer, J. (2004) 'Chapter3: Conquering the ECG', in *Cardiology Explained*. London: Remedica.

Aspiras, T.H. and Asari, V.K. (2011) 'Log power representation of EEG spectral bands for the recognition of emotional states of mind', *2011 8th International Conference on Information, Communications &amp; Signal Processing*[Preprint]. doi:10.1109/icics.2011.6174212.

Ayata, D., Yaslan, Y. and Kamasak, M. (2016) 'Emotion recognition via random forest and galvanic skin response: Comparison of time based feature sets, window sizes and wavelet approaches', *2016 Medical Technologies National Congress (TIPTEKNO)* [Preprint]. doi:10.1109/tiptekno.2016.7863130.

Azar, A.T. and El-Metwally, S.M. (2012) "Decision tree classifiers for Automated Medical Diagnosis," *Neural Computing and Applications*, 23(7-8), pp. 2387–2403. Available at: https://doi.org/10.1007/s00521-012-1196-7.

Aziz, H.A. (2018) 'Virtual reality programs applications in Healthcare', *Journal of Health &amp; Medical Informatics*, 09(01). doi:10.4172/2157-7420.1000305.

Azure.microsoft.com. 2020. *Facial Recognition | Microsoft Azure*. [online] Available at: <https://azure.microsoft.com/en-gb/services/cognitive-services/face/> [Accessed 24 June 2020].

Baker, F. (2023) *Statistical Press Notice NHS referral to treatment (RTT) waiting times data April 2023*, *NHS England*. Available at: https://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2023/06/Apr23-RTT-SPN-publication-version-PDF-427K.pdf (Accessed: 06 October 2023).

Balahur, A., Hermida, J. and Montoyo, A., 2012. Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems*, 53(4), pp.742-753.

Baños, R.M. *et al.* (2004) 'Immersion and emotion: Their impact on the sense of presence', *CyberPsychology &amp; Behavior*, 7(6), pp. 734–741. doi:10.1089/cpb.2004.7.734.

Beck, A., Sangoi, A., Leung, S., Marinelli, R., Nielsen, T., van de Vijver, M., West, R., van de Rijn, M. and Koller, D., 2011. Systematic Analysis of Breast Cancer Morphology Uncovers Stromal Features Associated with Survival. *Science Translational Medicine*, 3(108), pp.108ra113-108ra113.

Beheshti, A., Hashemi, V.M. and Wang, S. (2021) 'Towards predictive analytics in mental health care', *2021 International Joint Conference on Neural Networks (IJCNN)* [Preprint]. doi:10.1109/ijcnn52387.2021.9534233.

Bekelis, K. *et al.* (2017) 'Effect of an immersive preoperative virtual reality experience on patient reported outcomes', *Annals of Surgery*, 265(6), pp. 1068–1073. doi:10.1097/sla.0000000000002094.

Bergstra, J. *et al.* (2015) 'Hyperopt: A python library for model selection and hyperparameter optimization', *Computational Science &amp; Discovery*, 8(1), p. 014008. doi:10.1088/1749-4699/8/1/014008.

Berka, C., Levendowski, D.J., Lumicao, M.N., Yau, A., Davis, G., Zivkovic, V.T., Olmstead, R.E., Tremoulet, P.D. and Craven, P.L., (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine*, *78*(5), pp.B231-B244.

Bestpractice.bmj.com. 2020. *Depression In Adults - Monitoring | BMJ Best Practice*. [online] Available at: <https://bestpractice.bmj.com/topics/en-gb/55/monitoring> [Accessed 6 July 2020].

Bhardwaj, R., Nambiar, A. and Dutta, D. (2017). A Study of Machine Learning in Healthcare. *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*.

Bhuiyan, S.M., Adhami, R.R. and Khan, J.F. (2008) “A novel approach of fast and Adaptive Bidimensional Empirical Mode Decomposition,” *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*[Preprint]. Available at: https://doi.org/10.1109/icassp.2008.4517859.

Biau, G. and Scornet, E. (2016) ‘A Random Forest Guided Tour’, *TEST*, 25(2), pp. 197–227. doi:10.1007/s11749-016-0481-7.

Bickman, L. *et al.* (2011) "Effects of routine feedback to clinicians on mental health outcomes of youths: Results of a randomised trial," *Psychiatric Services*, 62(12), pp. 1423–1429. Available at: https://doi.org/10.1176/appi.ps.002052011.

Bin Rafiq, R. *et al.* (2020) ‘Validation methods to promote real-world applicability of machine learning in medicine’, *2020 3rd International Conference on Digital Medicine and Image Processing*, 12, pp. 13–19. doi:10.1145/3441369.3441372.

Biosemi (2001) *Biosemi EEG ECG EMG BSPM neuro amplifier electrodes*. Available at: https://www.biosemi.com/products.htm (Accessed: 05 August 2023).

Bisso, E. *et al.* (2020) ‘Immersive virtual reality applications in schizophrenia spectrum therapy: A systematic review’, *International Journal of Environmental Research and Public Health*, 17(17), p. 6111. doi:10.3390/ijerph17176111.

Bloomfield, P. and Ruiz de Villa, C. (2021) *Care Tech Landscape Review*. rep. Future Care Capital, pp. 17–17.

Bradley, M.M. and Lang, P.J. (1994) “Measuring emotion: The self-assessment manikin and the semantic differential,” *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1), pp. 49–59. Available at: https://doi.org/10.1016/0005-7916(94)90063-9.

Brockmyer, J.H. *et al.* (2009) ‘The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing’, *Journal of Experimental Social Psychology*, 45(4), pp. 624–634. doi:10.1016/j.jesp.2009.02.016.

Brown, T. *et al.* (2020) ‘Bringing virtual reality from clinical trials to clinical practice for the treatment of eating disorders: An example using virtual reality cue exposure therapy’, *Journal of Medical Internet Research*, 22(4). doi:10.2196/16386.

Bucci, S., Schwannauer, M. and Berry, N. (2019) ‘The Digital Revolution and its impact on Mental Health Care’, *Psychology and Psychotherapy: Theory, Research and Practice*, 92(2), pp. 277–297. doi:10.1111/papt.12222.

Cai, C.J. *et al.* (2019) ‘“hello ai”: Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making’, *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), pp. 1–24. doi:10.1145/3359206.

Cai, H. *et al.* (2021) ‘Combination of EOG and EEG for emotion recognition over different window sizes’, *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)* [Preprint]. doi:10.1109/ichms53169.2021.9582628.

Camurri, A., Mazzarino, B. and Volpe, G., 2004. Analysis of Expressive Gesture: The EyesWeb Expressive Gesture Processing Library. *Gesture-Based Communication in Human-Computer Interaction*, pp.460-467.

Canzian, L. and Musolesi, M. (2015) “Trajectories of depression,” *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15* [Preprint]. Available at: https://doi.org/10.1145/2750858.2805845.

Carpenter, C., Yang, C.-H. and West, D. (2021) "A comparison of sedentary behaviour as measured by the Fitbit and activpal in college students," *International Journal of Environmental Research and Public Health*, 18(8), p. 3914. Available at: https://doi.org/10.3390/ijerph18083914.

Casale, S., Russo, A., Scebba, G. and Serrano, S., 2008. Speech Emotion Classification Using Machine Learning Algorithms. *2008 IEEE International Conference on Semantic Computing*,.

Castellano, G., Kessous, L. and Caridakis, G., 2008. Emotion Recognition through Multiple Modalities: Face, Body Gesture, Speech. *Affect and Emotion in Human-Computer Interaction*, pp.92-103.

Center for Devices and Radiological Health (2020) *What is Digital Health?*, *U.S. Food and Drug Administration*. Available at: https://www.fda.gov/medical-devices/digital-health-center-excellence/what-digital-health#:~:text=The%20broad%20scope%20of%20digital,and%20telemedicine%2C%20and%20personalized%20medicine. (Accessed: 01 March 2024).

Chandwani, R., De, R. and Dwivedi, Y.K. (2018) ‘Telemedicine for low resource settings: Exploring the generative mechanisms’, *Technological Forecasting and Social Change*, 127, pp. 177–187. doi:10.1016/j.techfore.2017.06.014.

Chanel, G. *et al.* (2008) ‘Boredom, engagement and anxiety as indicators for adaptation to difficulty in games’, *Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era* [Preprint]. doi:10.1145/1457199.1457203.

Chanel, G. *et al.* (2011) ‘Emotion assessment from physiological signals for adaptation of game difficulty’, *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 41(6), pp. 1052–1063. doi:10.1109/tsmca.2011.2116000.

Char, D., Shah, N. and Magnus, D., 2018. Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *New England Journal of Medicine*, 378(11), pp.981-983.

Chatterjee, A., Gupta, U., Chinnakotla, M., Srikanth, R., Galley, M. and Agrawal, P., 2019. Understanding Emotions in Text Using Deep Learning and Big Data. *Computers in Human Behavior*, 93, pp.309-317.

Chen, Z., Lin, M., Chen, F., Lane, N., Cardone, G., Wang, R., Li, T., Chen, Y., Choudhury, T. and Cambell, A., 2013. Unobtrusive Sleep Monitoring using Smartphones. *Proceedings of the ICTs for improving Patients Rehabilitation Research Techniques*,.

Cheng, V.W. *et al.* (2019) ‘Gamification in apps and technologies for improving mental health and well-being: Systematic review’, *JMIR Mental Health*, 6(6). doi:10.2196/13717.

Chevance, G. *et al.* (2022) ‘Accuracy and precision of energy expenditure, heart rate, and steps measured by combined-sensing fitbits against reference measures: Systematic review and meta-analysis’, *JMIR mHealth and uHealth*, 10(4). doi:10.2196/35626.

Chiesa, V. *et al.* (2021) ‘Covid-19 pandemic: Health impact of staying at home, social distancing and “lockdown” measures—a systematic review of Systematic Reviews’, *Journal of Public Health*, 43(3). doi:10.1093/pubmed/fdab102.

Chikersal, P. *et al.* (2021) "Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing," *ACM Transactions on Computer-Human Interaction*, 28(1), pp. 1–41. Available at: https://doi.org/10.1145/3422821.

Cho, Y.M. *et al.* (2016) ‘A cross-sectional study of the association between mobile phone use and symptoms of ill health’, *Environmental Health and Toxicology*, 31. doi:10.5620/eht.e2016022.

Choi, E. et al. Doctor AI: predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare* 301–318 (2016).

Choudhury, T., Borriello, G., Consolvo, S., Haehnel, D., Harrison, B., Hemingway, B., Hightower, J., Klasnja, P., Koscher, K., LaMarca, A., Landay, J., LeGrand, L., Lester, J., Rahimi, A., Rea, A. and Wyatt, D., 2008. The Mobile Sensing Platform: An Embedded Activity Recognition System. *IEEE Pervasive Computing*, 7(2), pp.32-41.

Chowdhury, K.R., Sil, A. and Shukla, S.R. (2021) ‘Explaining a black-box sentiment analysis model with local interpretable model diagnostics explanation (lime)’, *Communications in Computer and Information Science*, pp. 90–101. doi:10.1007/978-3-030-81462-5_9.

Cicero, M., Bilbily, A., Colak, E., Dowdell, T., Gray, B., Perampaladas, K. and Barfett, J., 2017. Training and Validating a Deep Convolutional Neural Network for Computer-Aided Detection and Classification of Abnormalities on Frontal Chest Radiographs. *Investigative Radiology*, 52(5), pp.281-287.

Cleveland Clinic (2023) *Parasympathetic Nervous System (PSNS): What it is & function Cleveland Clinic*. Available at: https://my.clevelandclinic.org/health/body/23266-parasympathetic-nervous-system-psns (Accessed: March 10, 2023).

Cleveland Clinic (2023) *Sympathetic nervous system (SNS): What it is & function*, *Cleveland Clinic*. Available at: https://my.clevelandclinic.org/health/body/23262-sympathetic-nervous-system-sns-fight-or-flight (Accessed: March 10, 2023).

Clus, D. *et al.* (2018) ‘The use of virtual reality in patients with eating disorders: Systematic review’, *Journal of Medical Internet Research*, 20(4). doi:10.2196/jmir.7898.

Colomer Granero, A. *et al.* (2016) “A comparison of physiological signal analysis techniques and classifiers for automatic emotional evaluation of audiovisual contents,” *Frontiers in Computational Neuroscience*, 10. Available at: https://doi.org/10.3389/fncom.2016.00074.

Corno, L. and Mandinach, E.B. (1983) ‘The role of cognitive engagement in classroom learning and motivation’, *Educational Psychologist*, 18(2), pp. 88–108. doi:10.1080/00461528309529266.

Cortes, C. and Vapnik, V. (1995) ‘Support-Vector Networks’, *Machine Learning*, 20(3), pp. 273–297. doi:10.1007/bf00994018.

Cortiñas-Lorenzo, K. and Lacey, G. (2023) ‘Toward explainable affective computing: A Review’, *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–0. doi:10.1109/tnnls.2023.3270027.

Cowie, M.R. *et al.* (2016) ‘Electronic Health Records to facilitate clinical research’, *Clinical Research in Cardiology*, 106(1), pp. 1–9. doi:10.1007/s00392-016-1025-6.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York: Harper & Row.

Cuff, A. (2023) ‘The evolution of Digital Health and its continuing challenges’, *BMC Digital Health*, 1(1). doi:10.1186/s44247-022-00004-x.

Cuijpers, P. *et al.* (2019) ‘Effectiveness and acceptability of cognitive behavior therapy delivery formats in adults with depression’, *JAMA Psychiatry*, 76(7), p. 700. doi:10.1001/jamapsychiatry.2019.0268.

da Silva, F.L. (2009) ‘EEG: Origin and measurement’, *EEG - fMRI*, pp. 19–38. doi:10.1007/978-3-540-87919-0_2.

Dar, M.N. *et al.* (2020) ‘CNN and LSTM-based emotion charting using physiological signals’, *Sensors*, 20(16), p. 4551. doi:10.3390/s20164551.

Data acquisition, loggers, amplifiers, transducers, electrodes: BIOPAC (2020) *BIOPAC Systems, Inc.* Available at: https://www.biopac.com/ (Accessed: May 4, 2020).

Data and statistics (2022) *Centers for Disease Control and Prevention*. Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/sleep/data_statistics.html (Accessed: January 10, 2023).

Davenport, T. and Kalakota, R., 2019. The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, 6(2), pp.94-98.

De Angel, V. *et al.* (2022) ‘Digital Health Tools for the passive monitoring of depression: A systematic review of methods’, *npj Digital Medicine*, 5(1). doi:10.1038/s41746-021-00548-8.

de Carvalho, T., Noels, E., Wakkee, M., Udrea, A. and Nijsten, T., 2019. Development of Smartphone Apps for Skin Cancer Risk Assessment: Progress and Promise. *JMIR Dermatology*, 2(1), p.e13376.

De Choudhury M, Gamon M, Counts S, Horvitz E. 2013. Predicting depression via social media. Proc. 7th. Int. AAAI Conf. Weblogs Social Media, Boston, pp. 128–37. Palo Alto, CA: Assoc. Adv. Artif. Intell.

Dellazizzo, L. *et al.* (2020) ‘Evidence on virtual reality–based therapies for psychiatric disorders: Meta-review of Meta-analyses’, *Journal of Medical Internet Research*, 22(8). doi:10.2196/20889.

Dev, S. *et al.* (2022) ‘A predictive analytics approach for stroke prediction using machine learning and Neural Networks’, *Healthcare Analytics*, 2, p. 100032. doi:10.1016/j.health.2022.100032.

Difede, J. and Hoffman, H.G. (2002) “Virtual reality exposure therapy for world trade center post-traumatic stress disorder: A case report,” *CyberPsychology & Behavior*, 5(6), pp. 529–535. Available at: https://doi.org/10.1089/109493102321018169.

Digital Health (2023) *World Health Organization*. Available at: https://www.who.int/health-topics/digital-health#tab=tab_3 (Accessed: 14 December 2023).

Diveica, V. *et al.* (2023) ‘Graded functional organization in the left inferior frontal gyrus: Evidence from task-free and task-based functional connectivity’, *Cerebral Cortex*, 33(23), pp. 11384–11399. doi:10.1093/cercor/bhad373.

Diykh, M., Li, Y. and Wen, P. (2016) ‘EEG sleep stages classification based on time domain features and structural graph similarity’, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 24(11), pp. 1159–1168. doi:10.1109/tnsre.2016.2552539.

Domínguez-Jiménez, J.A. *et al.* (2020) “A machine learning model for emotion recognition from physiological signals,” *Biomedical Signal Processing and Control*, 55, p. 101646. Available at: https://doi.org/10.1016/j.bspc.2019.101646.

Don H. Hockenbury and Sandra E. Hockenbury. 2010. Discovering psychology and study guide. Macmillan.

Dong, Y. *et al.* (2022) “Detection of arrhythmia in 12-lead varied-length ECG using multi-branch Signal Fusion Network,” *Physiological Measurement*, 43(10), p. 105009. Available at: https://doi.org/10.1088/1361-6579/ac7938.

Dozois, D.J. *et al.* (2014) "The CPA presidential task force on evidence-based practice of psychological treatments.," *Canadian Psychology / Psychologie canadienne*, 55(3), pp. 153–160. Available at: https://doi.org/10.1037/a0035767.

Drissi, N. *et al.* (2021) ‘A systematic literature review on e-mental health solutions to assist health care workers during COVID-19’, *Telemedicine and e-Health*, 27(6), pp. 594–602. doi:10.1089/tmj.2020.0287.

du Sert, O.P. *et al.* (2018) ‘Virtual reality therapy for refractory auditory verbal hallucinations in schizophrenia: A pilot clinical trial’, *Schizophrenia Research*, 197, pp. 176–181. doi:10.1016/j.schres.2018.02.031.

Ebrahimi, Z. *et al.* (2020) “A review on deep learning methods for ECG Arrhythmia Classification,” *Expert Systems with Applications: X*, 7, p. 100033. Available at: https://doi.org/10.1016/j.eswax.2020.100033.

Egger, M., Ley, M. and Hanke, S., 2019. Emotion Recognition from Physiological Signal Analysis: A Review. Electronic Notes in Theoretical Computer Science, 343, pp.35-55.

Ekiz, D. *et al.* (2021) ‘End-to-end deep multi-modal physiological authentication with Smartbands’, *IEEE Sensors Journal*, 21(13), pp. 14977–14986. doi:10.1109/jsen.2021.3073888.

Ekman, P. and Friesen, W., (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2), pp.124-129.

Ekman, P., (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4), pp.169-200.

Electroencephalogram (EEG) (2022) *NHS choices*. Available at: https://www.nhs.uk/conditions/electroencephalogram/ (Accessed: 10 August 2023).

Eshuis, L.V. *et al.* (2021) ‘Efficacy of immersive PTSD treatments: A systematic review of virtual and augmented reality exposure therapy and a meta-analysis of virtual reality exposure therapy’, *Journal of Psychiatric Research*, 143, pp. 516–527. doi:10.1016/j.jpsychires.2020.11.030.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S. and Dean, J., (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), pp.24-29. doi: https://doi.org/10.1038/s41591-018-0316-z.

Evans, R.S. (2016) ‘Electronic health records: Then, now, and in the future’, *Yearbook of Medical Informatics*, 25(S 01). doi:10.15265/iys-2016-s006.

Eyes on speech communication (2023) *Vokaturi*. Available at: https://vokaturi.com/ (Accessed: 29 February 2024).

Falconer, C.J. *et al.* (2016) ‘Embodying self-compassion within virtual reality and its effects on patients with depression’, *BJPsych Open*, 2(1), pp. 74–80. doi:10.1192/bjpo.bp.115.002147.

Farhan, A.A. *et al.* (2016) "Behavior vs. introspection: Refining prediction of clinical depression via smartphone sensing data," *2016 IEEE Wireless Health (WH)* [Preprint]. Available at: https://doi.org/10.1109/wh.2016.7764553.

Farias, F.A. *et al.* (2020) ‘Remote Patient Monitoring: A systematic review’, *Telemedicine and e-Health*, 26(5), pp. 576–583. doi:10.1089/tmj.2019.0066.

Faust, O. *et al.* (2018) “Deep learning for healthcare applications based on Physiological Signals: A Review,” *Computer Methods and Programs in Biomedicine*, 161, pp. 1–13. Available at: https://doi.org/10.1016/j.cmpb.2018.04.005.

Feradov, F. and Ganchev, T., 2015. Ranking of EEG time-domain features on the negative emotions recognition task. *Annual journal of Electronics*, *9*, pp.26-29.

Field, M. (1996) *Telemedicine: A Guide to Assessing Telecommunications in Health Care.* rep. Washington DC: National Academies Press (US).

Field, T.A., Beeson, E.T. and Jones, L.K. (2015) ‘The new abcs: A practitioner’s guide to neuroscience-informed cognitive-behavior therapy’, *Journal of Mental Health Counseling*, 37(3), pp. 206–220. doi:10.17744/1040-2861-37.3.206.

Firth, J. *et al.* (2015) ‘Mobile phone ownership and endorsement of “mhealth” among people with psychosis: A meta-analysis of cross-sectional studies’, *Schizophrenia Bulletin*, 42(2), pp. 448–455. doi:10.1093/schbul/sbv132.

Fitbit help (2022) *Fitbit MyHelp*. Available at: https://help.fitbit.com/articles/en_US/Help_article/2163.htm (Accessed: February 17, 2022).

Fleming, T., Poppelaars, M. and Thabrew, H. (2023) ‘The role of gamification in Digital Mental Health’, *World Psychiatry*, 22(1), pp. 46–47. doi:10.1002/wps.21041.

Freeman, D. *et al.* (2016) ‘Virtual reality in the treatment of persecutory delusions: randomised controlled experimental study testing how to reduce delusional conviction’, *British Journal of Psychiatry*, 209(1), pp. 62–67. doi:10.1192/bjp.bp.115.176438.

Freeman, D. *et al.* (2022) ‘Virtual reality (VR) therapy for patients with psychosis: Satisfaction and side effects’, *Psychological Medicine*, 53(10), pp. 4373–4384. doi:10.1017/s0033291722001167.

Gábana Arellano, D., Tokarchuk, L. and Gunes, H. (2016) ‘Measuring affective, physiological and behavioural differences in solo, competitive and collaborative games’, *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pp. 184–193. doi:10.1007/978-3-319-49616-0_18.

Garcia-Garcia, J., Penichet, V. and Lozano, M., 2017. Emotion detection. Proceedings of the XVIII International Conference on Human Computer Interaction - Interacción ’17,.

Gaston, L. (1998) ‘Alliance, technique, and their interactions in predicting outcome of behavioral, cognitive, and brief dynamic therapy’, *Psychotherapy Research*, 8(2), pp. 190–209. doi:10.1093/ptr/8.2.190.

Gega, L. *et al.* (2022) ‘Digital Interventions in mental health: Evidence syntheses and economic modelling’, *Health Technology Assessment*, 26(1), pp. 1–182. doi:10.3310/rcti6942.

Geraets, C.N.W. *et al.* (2021) 'Advances in immersive virtual reality interventions for mental disorders: A new reality?', *Current Opinion in Psychology*, 41, pp. 40–45. doi:10.1016/j.copsyc.2021.02.004.

Gerlings, J., Jensen, M.S. and Shollo, A. (2021) ‘Explainable AI, but explainable to whom? an exploratory case study of XAI in Healthcare’, *Handbook of Artificial Intelligence in Healthcare*, pp. 169–198. doi:10.1007/978-3-030-83620-7_7.

Geurts, P., Ernst, D. and Wehenkel, L. (2006) ‘Extremely randomized trees’, *Machine Learning*, 63(1), pp. 3–42. doi:10.1007/s10994-006-6226-1.

Ghandeharioun, A. *et al.* (2017) ‘Objective assessment of depressive symptoms with machine learning and wearable sensors data’, *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*[Preprint]. doi:10.1109/acii.2017.8273620.

Gitas, G. *et al.* (2022) ‘Robotic Surgery in gynecology: Is the future already here?’, *Minimally Invasive Therapy &amp; Allied Technologies*, 31(6), pp. 815–824. doi:10.1080/13645706.2021.2010763.

Gleacher, A.A. *et al.* (2015) "Implementing a measurement feedback system in Community Mental Health Clinics: A case study of multilevel barriers and facilitators," *Administration and Policy in Mental Health and Mental Health Services Research*, 43(3), pp. 426–440. Available at: https://doi.org/10.1007/s10488-015-0642-0.

Gomes-Schwartz, B. (1978) ‘Effective ingredients in psychotherapy: Prediction of outcome from process variables.’, *Journal of Consulting and Clinical Psychology*, 46(5), pp. 1023–1035. doi:10.1037/0022-006x.46.5.1023.

Goodfellow, S.D., Goodwin, A., Greer, R., Laussen, P.C., Mazwi, M. and Eytan, D., 2018, November. Towards understanding ECG rhythm classification using convolutional neural networks and attention mappings. In *Machine learning for healthcare conference* (pp. 83-101). PMLR.

Gouverneur, P. *et al.* (2023) ‘Explainable artificial intelligence (XAI) in pain research: Understanding the role of electrodermal activity for Automated Pain Recognition’, *Sensors*, 23(4), p. 1959. doi:10.3390/s23041959.

Gramfort, A. (2013) ‘Meg and EEG data analysis with MNE-Python’, *Frontiers in Neuroscience*, 7. doi:10.3389/fnins.2013.00267.

Granato, M. *et al.* (2018) ‘Feature extraction and selection for real-time emotion recognition in video games players’, *2018 14th International Conference on Signal-Image Technology &amp; Internet-Based Systems (SITIS)* [Preprint]. doi:10.1109/sitis.2018.00115.

Greenberg, L.S. (2004) 'Emotion–focused therapy', *Clinical Psychology &amp; Psychotherapy*, 11(1), pp. 3–16. doi:10.1002/cpp.388.

Griffiths, C. *et al.* (2022) "Investigation of physical activity, sleep, and mental health recovery in treatment resistant depression (TRD) patients receiving repetitive transcranial magnetic stimulation (rtms) treatment," *Journal of Affective Disorders Reports*, 8, p. 100337. Available at: https://doi.org/10.1016/j.jadr.2022.100337.

Grigorovici, D.M. and Constantin, C.D. (2004) "Experiencing interactive advertising beyond Rich Media," *Journal of Interactive Advertising*, 5(1), pp. 22–36. Available at: https://doi.org/10.1080/15252019.2004.10722091.

Guo, C. and Chen, J. (2023) 'Big Data Analytics in Healthcare', *Translational Systems Sciences*, pp. 27–70. doi:10.1007/978-981-99-1075-5_2.

Gutner, C.A. *et al.* (2016) 'Does timing matter? examining the impact of session timing on outcome.', *Journal of Consulting and Clinical Psychology*, 84(12), pp. 1108–1115. doi:10.1037/ccp0000120.

Hafeez, T. *et al.* (2021) 'EEG in Game user analysis: A framework for expertise classification during gameplay', *PLOS ONE*, 16(6). doi:10.1371/journal.pone.0246913.

Haghayegh, S. *et al.* (2019) 'Accuracy of wristband fitbit models in assessing sleep: Systematic review and meta-analysis', *Journal of Medical Internet Research*, 21(11). doi:10.2196/16273.

Halbig, A. *et al.* (2022) 'Opportunities and challenges of virtual reality in healthcare – a domain experts inquiry', *Frontiers in Virtual Reality*, 3. doi:10.3389/frvir.2022.837616.

Hamdi, NR *et al.* (2021) "APA GUIDELINES on Evidence-Based Psychological Practice in Health Care." Washington DC: https://www.apa.org/about/policy/psychological-practice-health-care.pdf.

Han, S.Y. *et al.* (2020) 'Diagnostic prediction model development using data from dried blood spot proteomics and a digital mental health assessment to identify major depressive disorder among individuals presenting with low mood', *Brain, Behavior, and Immunity*, 90, pp. 184–195. doi:10.1016/j.bbi.2020.08.011.

Hanna, M.G. *et al.* (2018) 'Augmented Reality Technology using Microsoft HoloLens in anatomic pathology', *Archives of Pathology &amp; Laboratory Medicine*, 142(5), pp. 638–644. doi:10.5858/arpa.2017-0189-oa.

Hassouneh, A., Mutawa, A.M. and Murugappan, M. (2020) “Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods,” *Informatics in Medicine Unlocked*, 20, p. 100372. Available at: https://doi.org/10.1016/j.imu.2020.100372.

Hastie T, Tibshirani R: Generalized Additive Models. 1990, London: Chapman & Hall

Hastie, T., Tibshirani, R. and Friedman, J.H. (2001) *The elements of statistical learning data mining, Inference, and prediction*. New York, NY: Springer New York.

Hawker, G.A. *et al.* (2011) “Measures of adult pain: Visual Analog Scale for pain (Vas Pain), numeric rating scale for pain (NRS Pain), McGill Pain Questionnaire (MPQ), short-form mcgill pain questionnaire (SF-MPQ), chronic pain grade scale (CPGS), short form-36 bodily pain scale (SF,” *Arthritis Care & Research*, 63(S11). Available at: https://doi.org/10.1002/acr.20543.

Heimerl, A. *et al.* (2022) ‘Unraveling ML models of emotion with nova: Multi-level explainable AI for Non-Experts’, *IEEE Transactions on Affective Computing*, 13(3), pp. 1155–1167. doi:10.1109/taffc.2020.3043603.

Higuera-Trujillo, J.L., López-Tarruella Maldonado, J. and Llinares Millán, C. (2017) ‘Psychological and physiological human responses to simulated and Real Environments: A comparison between photographs, 360° panoramas, and virtual reality’, *Applied Ergonomics*, 65, pp. 398–409. doi:10.1016/j.apergo.2017.05.006.

Holzinger, A., Biemann, C., Pattichis, C.S. and Kell, D.B., 2017. What do we need to build explainable AI systems for the medical domain?. *arXiv preprint arXiv:1712.09923*.

Hoque, R. and DelRosso, L.M. (2014) ‘Epileptiform discharges during slow wave sleep on polysomnogram’, *Journal of Clinical Sleep Medicine*, 10(03), pp. 336–339. doi:10.5664/jcsm.3546.

Horvath, A.O. *et al.* (2011) ‘Alliance in individual psychotherapy’, *Psychotherapy Relationships That Work*, pp. 25–69. doi:10.1093/acprof:oso/9780199737208.003.0002.

Hosseinifard, B., Moradi, M. and Rostami, R., 2013. Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal. *Computer Methods and Programs in Biomedicine*, 109(3), pp.339-345.

House of Commons *et al.* (2023) *Progress in improving NHS mental health services*. House of Commons.

Huynh, S. *et al.* (2018) ‘Engagemon’, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1), pp. 1–27. doi:10.1145/3191745.

Iffland, B. *et al.* (2019) “Cardiac reactions to emotional words in adolescents and young adults with PTSD after child abuse,” *Psychophysiology*, 57(1). Available at: https://doi.org/10.1111/psyp.13470.

Intarasirisawat, J. *et al.* (2020) "An automated mobile game-based screening tool for patients with alcohol dependence," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3), pp. 1–23. Available at: https://doi.org/10.1145/3411837.

Ionita, G. and Fitzpatrick, M. (2014) "Bringing science to clinical practice: A Canadian survey of psychological practice and usage of progress monitoring measures.," *Canadian Psychology / Psychologie canadienne*, 55(3), pp. 187–196. Available at: https://doi.org/10.1037/a0037355.

Ismail Fawaz, H. *et al.* (2019) ‘Deep Learning for Time Series classification: A Review’, *Data Mining and Knowledge Discovery*, 33(4), pp. 917–963. doi:10.1007/s10618-019-00619-1.

Iwagami, H. *et al.* (2020) ‘Artificial Intelligence for the detection of esophageal and esophagogastric junctional adenocarcinoma’, *Journal of Gastroenterology and Hepatology*, 36(1), pp. 131–136. doi:10.1111/jgh.15136.

J. Devoe, D. *et al.* (2022) ‘The impact of the covid-19 pandemic on Eating disorders: A systematic review’, *International Journal of Eating Disorders*, 56(1), pp. 5–25. doi:10.1002/eat.23704.

Jaiswal, S. and Nandi, G.C. (2019) ‘Robust real-time emotion detection system using CNN architecture’, *Neural Computing and Applications*, 32(15), pp. 11253–11262. doi:10.1007/s00521-019-04564-4.

Jalilifard, A., Pizzolato, E.B. and Islam, M.K. (2016) “Emotion classification using single-channel SCALP-EEG recording,” *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* [Preprint]. Available at: https://doi.org/10.1109/embc.2016.7590833.

Janiesch, C., Zschech, P. and Heinrich, K. (2021) ‘Machine learning and deep learning’, *Electronic Markets*, 31(3), pp. 685–695. doi:10.1007/s12525-021-00475-2.

Javaid, M. and Haleem, A. (2020) ‘Virtual reality applications toward medical field’, *Clinical Epidemiology and Global Health*, 8(2), pp. 600–605. doi:10.1016/j.cegh.2019.12.010.

Jensen-Doss, A. *et al.* (2016) ‘Monitoring treatment progress and providing feedback is viewed favorably but rarely used in practice’, *Administration and Policy in Mental Health and Mental Health Services Research*, 45(1), pp. 48–61. doi:10.1007/s10488-016-0763-0.

Jerritta, S. *et al.* (2011) ‘Physiological signals based human emotion recognition: A Review’, *2011 IEEE 7th International Colloquium on Signal Processing and its Applications* [Preprint]. doi:10.1109/cspa.2011.5759912.

Jones, T., Moore, T. and Choo, J. (2016) ‘The impact of virtual reality on Chronic pain’, *PLOS ONE*, 11(12). doi:10.1371/journal.pone.0167523.

Joshi, J., Goecke, R., Alghowinem, S., Dhall, A., Wagner, M., Epps, J., Parker, G. and Breakspear, M., 2013. Multimodal assistive technologies for depression diagnosis and monitoring. *Journal on Multimodal User Interfaces*, 7(3), pp.217-228.

Joy, T.T. *et al.* (2016) ‘Hyperparameter tuning for big data using Bayesian optimisation’, *2016 23rd International Conference on Pattern Recognition (ICPR)* [Preprint]. doi:10.1109/icpr.2016.7900023.

Kamran Ul haq, A. *et al.* (2020) ‘Data Analytics in mental healthcare’, *Scientific Programming*, 2020, pp. 1–9. doi:10.1155/2020/2024160.

Kanjo, E., Younis, E.M.G. and Ang, C.S. (2019) ‘Deep Learning Analysis of Mobile Physiological, environmental and location sensor data for emotion detection’, *Information Fusion*, 49, pp. 46–56. doi:10.1016/j.inffus.2018.09.001.

Kappelman, L.A. (1995) ‘Measuring user involvement’, *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, 26(2–3), pp. 65–86. doi:10.1145/217278.217286.

Kassahun, Y., Yu, B., Tibebu, A.T. *et al.* Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions. *Int J CARS* 11**,** 553–568 (2016).

Katsigiannis, S. and Ramzan, N. (2018) “Dreamer: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices,” *IEEE Journal of Biomedical and Health Informatics*, 22(1), pp. 98–107. Available at: https://doi.org/10.1109/jbhi.2017.2688239.

Katsis, C., Katertsidis, N., Ganiatsas, G. and Fotiadis, D., 2008. Toward Emotion Recognition in Car-Racing Drivers: A Biosignal Processing Approach. *IEEE Transactions*

Keane, M.T. and Kenny, E.M. (2019) ‘How case-based reasoning explains neural networks: A theoretical analysis of XAI using Post-Hoc Explanation-by-example from a survey of Ann-CBR twin-systems’, *Case-Based Reasoning Research and Development*, pp. 155–171. doi:10.1007/978-3-030-29249-2_11.

Ketkar, N. (2017) ‘Introduction to keras’, *Deep Learning with Python*, pp. 97–111. doi:10.1007/978-1-4842-2766-4_7.

Khalane, A. *et al.* (2023) ‘Evaluating significant features in context-aware multimodal emotion recognition with xai methods’, *Expert Systems* [Preprint]. doi:10.1111/exsy.13403.

Khanal, S., Barroso, J., Lopes, N., Sampaio, J. and Filipe, V., 2018. Performance analysis of Microsoft's and Google's Emotion Recognition API using pose-invariant faces. *Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion - DSAI 2018*,.

Kim, J., Campbell, A., de Ávila, B. and Wang, J., 2019. Wearable biosensors for healthcare monitoring. *Nature Biotechnology*, 37(4), pp.389-406.

Kiper, P. *et al.* (2022) ‘Effects of immersive virtual therapy as a method supporting recovery of depressive symptoms in post-stroke rehabilitation: Randomized Controlled Trial’, *Clinical Interventions in Aging*, Volume 17, pp. 1673–1685. doi:10.2147/cia.s375754.

Kılıç, A.C., Karakuş, A. and Alptekin, E. (2022) ‘Prediction of university students’ subjective well-being with sleep and physical activity data using classification algorithms’, *Procedia Computer Science*, 207, pp. 2648–2657. doi:10.1016/j.procs.2022.09.323.

Koelstra, S. *et al.* (2012) “DEAP: A database for emotion analysis ;using physiological signals,” *IEEE Transactions on Affective Computing*, 3(1), pp. 18–31. Available at: https://doi.org/10.1109/t-affc.2011.15.

Koochaki, F. and Najafizadeh, L. (2019) ‘Eye gaze-based early intent prediction utilizing CNN-LSTM’, *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* [Preprint]. doi:10.1109/embc.2019.8857054.

Kosunen, I. *et al.* (2016) “RelaWorld,” *Proceedings of the 21st International Conference on Intelligent User Interfaces*[Preprint]. Available at: https://doi.org/10.1145/2856767.2856796.

Kotte, A. *et al.* (2016) "Facilitators and barriers of implementing a measurement feedback system in public youth mental health," *Administration and Policy in Mental Health and Mental Health Services Research*, 43(6), pp. 861–878. Available at: https://doi.org/10.1007/s10488-016-0729-2.

Kroenke, K., Spitzer, R.L. and Williams, J.B. (2001) "The PHQ-9," *Journal of General Internal Medicine*, 16(9), pp. 606–613. Available at: https://doi.org/10.1046/j.1525-1497.2001.016009606.x.

Kumar, M.A., Vimala, R. and Britto, K.R.A. (2019) “A cognitive technology based healthcare monitoring system and medical data transmission,” *Measurement*, 146, pp. 322–332. Available at: https://doi.org/10.1016/j.measurement.2019.03.017.

Kumari, M. *et al.* (2009) "Self-reported sleep duration and sleep disturbance are independently associated with cortisol secretion in the Whitehall II Study," *The Journal of Clinical Endocrinology & Metabolism*, 94(12), pp. 4801–4809. Available at: https://doi.org/10.1210/jc.2009-0555.

La Ferla, M. (2023) ‘An XAI approach to deep learning models in the detection of DCIS’, *IFIP Advances in Information and Communication Technology*, pp. 409–420. doi:10.1007/978-3-031-34171-7_33.

La Paglia, F., La Cascia, C., Rizzo, R., Sanna, M., Cangialosi, F., Sideli, L., Francomano, A., Riva, G. and La Barbera, D., (2016). Virtual reality environments to rehabilitation attention deficits in schizophrenic patients. *CYBER THERAPY AND REHABILITATION MAGAZINE*, *9*(1), pp.34-34.

Lake, J., (2017). Urgent Need for Improved Mental Health Care and a More Collaborative Model of Care. *The Permanente Journal*,.

Lambert, M.J. *et al.* (2003) "Is it time for clinicians to routinely track patient outcome? A meta-analysis.," *Clinical Psychology: Science and Practice*, 10(3), pp. 288–301. Available at: https://doi.org/10.1093/clipsy.bpg025.

Lambert, M.J., Whipple, J.L. and Kleinstäuber, M. (2018) "Collecting and delivering progress feedback: A meta-analysis of routine outcome monitoring.," *Psychotherapy*, 55(4), pp. 520–537. Available at: https://doi.org/10.1037/pst0000167.

Lance, B. and Marsella, S.C. (2008) "The relation between gaze behavior and the attribution of emotion: An empirical study," *Intelligent Virtual Agents*, pp. 1–14. Available at: https://doi.org/10.1007/978-3-540-85483-8_1.

Landau, H.J. (1967) 'Sampling, data transmission, and the Nyquist rate', *Proceedings of the IEEE*, 55(10), pp. 1701–1706. doi:10.1109/proc.1967.5962.

Lane, H.C. and D'Mello, S.K. (2018) "Uses of physiological monitoring in Intelligent Learning Environments: A review of research, evidence, and technologies," *Mind, Brain and Technology*, pp. 67–86. Available at: https://doi.org/10.1007/978-3-030-02631-8_5.

Lecun, Y. *et al.* (1998) "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 86(11), pp. 2278–2324. Available at: https://doi.org/10.1109/5.726791.

Lee, J., Byun, W., Keill, A., Dinkel, D. and Seo, Y., 2018. Comparison of Wearable Trackers' Ability to Estimate Sleep. *International Journal of Environmental Research and Public Health*, 15(6), p.1265.

Lee, S. *et al.* (2020) "Using physiological recordings for studying user experience: Case of conversational agent-equipped TV," *International Journal of Human–Computer Interaction*, 36(9), pp. 815–827. Available at: https://doi.org/10.1080/10447318.2019.1693166.

Li, A. *et al.* (2022) 'MNE-ICALabel: Automatically annotating ICA componentswith ICLabel in python', *Journal of Open Source Software*, 7(76), p. 4484. doi:10.21105/joss.04484.

Li, J. (2023)' Digital Technologies for mental health improvements in the COVID-19 pandemic: A scoping review', *BMC Public Health*, 23(1). doi:10.1186/s12889-023-15302-w.

Li, Z. *et al.* (2020) ‘Demystifying Signal Processing Techniques to extract resting-state EEG features for psychologists’, *Brain Science Advances*, 6(3), pp. 189–209. doi:10.26599/bsa.2020.9050019.

Li, Z. *et al.* (2022) ‘A survey of Convolutional Neural Networks: Analysis, applications, and prospects’, *IEEE Transactions on Neural Networks and Learning Systems*, 33(12), pp. 6999–7019. doi:10.1109/tnnls.2021.3084827.

Liao, J. *et al.* (2021) ‘Recognizing diseases with multivariate physiological signals by a deepcnn-LSTM network’, *Applied Intelligence*, 51(11), pp. 7933–7945. doi:10.1007/s10489-021-02309-2.

Liew, W.S., Loo, C.K. and Wermter, S. (2021) ‘Emotion recognition using explainable genetically optimized fuzzy art ensembles’, *IEEE Access*, 9, pp. 61513–61531. doi:10.1109/access.2021.3072120.

LikamWa, R., Liu, Y., Lane, N. and Zhong, L., 2013. MoodScope. Proceeding of the 11th annual international conference on Mobile systems, applications, and services - MobiSys ’13,.

Lim, J.Z., Mountstephens, J. and Teo, J. (2020) “Emotion recognition using eye-tracking: Taxonomy, review and current challenges,” *Sensors*, 20(8), p. 2384. Available at: https://doi.org/10.3390/s20082384.

Lin, T. and Imamiya, A. (2006) “Evaluating usability based on multimodal information,” *Proceedings of the 8th international conference on Multimodal interfaces* [Preprint]. Available at: https://doi.org/10.1145/1180995.1181063.

Lin, T. *et al.* (2022) ‘A survey of Transformers’, *AI Open*, 3, pp. 111–132. doi:10.1016/j.aiopen.2022.10.001.

Lindner, P. *et al.* (2019) ‘Attitudes toward and familiarity with virtual reality therapy among practicing Cognitive Behavior Therapists: A cross-sectional survey study in the era of consumer VR platforms’, *Frontiers in Psychology*, 10. doi:10.3389/fpsyg.2019.00176.

Lindner, P. *et al.* (2020) ‘Virtual reality exposure therapy for public speaking anxiety in routine care: A single-subject effectiveness trial’, *Cognitive Behaviour Therapy*, 50(1), pp. 67–87. doi:10.1080/16506073.2020.1795240.

Liu, Z. *et al.* (2022) ‘Virtual reality aided therapy towards Health 4.0: A two-decade bibliometric analysis’, *International Journal of Environmental Research and Public Health*, 19(3), p. 1525. doi:10.3390/ijerph19031525.

Loh, H.W. *et al.* (2022) ‘Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)’, *Computer Methods and Programs in Biomedicine*, 226, p. 107161. doi:10.1016/j.cmpb.2022.107161.

Loh, H.W. *et al.* (2023) ‘Deep neural network technique for automated detection of ADHD and CD using ECG Signal’, *Computer Methods and Programs in Biomedicine*, 241, p. 107775. doi:10.1016/j.cmpb.2023.107775.

Lu, J. *et al.* (2018) ‘Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning’, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1), pp. 1–21. doi:10.1145/3191753.

Lu, L. *et al.* (2020) ‘Wearable Health Devices in health care: Narrative systematic review’, *JMIR mHealth and uHealth*, 8(11). doi:10.2196/18907.

Lu, Y., Zheng, W.L., Li, B. and Lu, B.L., (2015). Combining Eye Movements and EEG to Enhance Emotion Recognition. In *IJCAI* (Vol. 15, pp. 1170-1176).

Luck, SJ (2014) *An introduction to the event-related potential technique*. Cambridge (Mass.): MIT Press.

Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Lutz, W. *et al.* (2021) 'Measuring, Predicting and Tracking Change in Psychotherapy', in *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change*. John Wiley & Sons, pp. 89–119.

Luxton, D., Anderson, S. and Anderson, M., 2016. Ethical Issues and Artificial Intelligence Technologies in Behavioral and Mental Health Care. *Artificial Intelligence in Behavioral and Mental Health Care*, pp.255-276.

M., G. *et al.* (2023) 'Explainable deep learning-based approach for Multilabel classification of Electrocardiogram', *IEEE Transactions on Engineering Management*, 70(8), pp. 2787–2799. doi:10.1109/tem.2021.3104751.

Macedonio, M.F. *et al.* (2007) "Immersiveness and physiological arousal within panoramic video-based virtual reality," *CyberPsychology & Behavior*, 10(4), pp. 508–515. Available at: https://doi.org/10.1089/cpb.2007.9997.

Madan, A., Cebrian, M., Lazer, D. and Pentland, A., 2010. Social sensing for epidemiological behavior change. *Proceedings of the 12th ACM international conference on Ubiquitous computing*,.

Mahmood, M. *et al.* (2019) "Fully portable and wireless universal brain–machine interfaces enabled by Flexible Scalp Electronics and deep learning algorithm," *Nature Machine Intelligence*, 1(9), pp. 412–422. Available at: https://doi.org/10.1038/s42256-019-0091-7.

Maximising the potential of digital in mental health. rep. (2023) Available at: https://www.nhsconfed.org/publications/maximising-potential-digital-mental-health#:~:text=before%20conditions%20worsen.-,Benefits,overall%20mental%20wellbeing%20for%20users (Accessed: 2023).

Mayor Torres, J.M. *et al.* (2023) 'Evaluation of interpretability for deep learning algorithms in EEG EMOTION RECOGNITION: A case study in autism', *Artificial Intelligence in Medicine*, 143, p. 102545. doi:10.1016/j.artmed.2023.102545.

McDaid, D. *et al.* (2022) *The economic case for investing in the prevention of mental health conditions in the UK (Summary)*. rep.

McGinnis, R., McGinnis, E., Hruschak, J., Lopez-Duran, N., Fitzgerald, K., Rosenblum, K. and Muzik, M., 2018. Rapid Anxiety and Depression Diagnosis in Young Children Enabled by Wearable Sensors and Machine Learning. *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*,.

McKinney, S., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C., King, D., Ledsam, J., Melnick, D., Mostofi, H., Peng, L., Reicher, J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K., De Fauw, J. and Shetty, S., 2020. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), pp.89-94.

McManus S, Bebbington P, Jenkins R, Brugha T. (eds.) (2016). Mental health and wellbeing in England: Adult psychiatric morbidity survey 2014.

McManus, S., Meltzer, H., Brugha, T. S., Bebbington, P. E., & Jenkins, R. (2009). Adult psychiatric morbidity in England, 2007: results of a household survey.

Meditation and sleep made simple (2024) *Headspace*. Available at: https://www.headspace.com/ (Accessed: 04 March 2024).

Meerlo, P., Sgoifo, A. and Suchecki, D. (2008) "Restricted and disrupted sleep: Effects on autonomic function, neuroendocrine stress systems and stress responsivity," *Sleep Medicine Reviews*, 12(3), pp. 197–210. Available at: https://doi.org/10.1016/j.smrv.2007.07.007.

Mehrotra, A. and Musolesi, M. (2018) "Using autoencoders to automatically extract mobility features for predicting depressive states," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3), pp. 1–20. Available at: https://doi.org/10.1145/3264937.

Mehta, H. and Passi, K. (2022) 'Social media hate speech detection using Explainable Artificial Intelligence (XAI)', *Algorithms*, 15(8), p. 291. doi:10.3390/a15080291.

Mendivil, A., Holloway, R. and Boggess, J., 2009. Emergence of robotic assisted surgery in gynecologic oncology: American perspective. *Gynecologic Oncology*, 114(2), pp.S24-S31.

Mental health (2023) *World Health Organization*. Available at: https://www.who.int/health-topics/mental-health#tab=tab_1 (Accessed: 14 December 2023).

Meta (2023) *Meta quest 2: Immersive all-in-one VR headset: Meta store*, *Meta*. Available at: https://www.meta.com/gb/en/quest/products/quest-2/?utm_source=www.oculus.com&utm_medium=oculusredirect (Accessed: 02 February 2023).

Mindtools.com. 2020. *Body Language: Picking Up And Understanding Nonverbal Signals*. [online] Available at: <https://www.mindtools.com/pages/article/Body_Language.htm> [Accessed 24 June 2020].

Miotto, R., Li, L., Kidd, B. and Dudley, J., 2016. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*, 6(1).

Mishra, S. *et al.* (2020) “Soft, wireless periocular wearable electronics for real-time detection of eye vergence in a virtual reality toward Mobile Eye Therapies,” *Science Advances*, 6(11). Available at: https://doi.org/10.1126/sciadv.aay1729.

Mohd, N.S. *et al.* (2019) “Millennial tourist emotional experience in technological engagement at destination,” *International Journal of Built Environment and Sustainability*, 6(1-2), pp. 129–135. Available at: https://doi.org/10.11113/ijbes.v6.n1-2.396.

Mohr, D., Zhang, M. and Schueller, S., 2017. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annual Review of Clinical Psychology*, 13(1), pp.23-47.

Monteiro, D. *et al.* (2018) ‘Evaluating engagement of virtual reality games based on first and third person perspective using EEG and subjective metrics’, *2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)* [Preprint]. doi:10.1109/aivr.2018.00015.

Mou, L. *et al.* (2021) ‘Driver stress detection via multimodal fusion using attention-based CNN-LSTM’, *Expert Systems with Applications*, 173, p. 114693. doi:10.1016/j.eswa.2021.114693.

Murphy, J.K. *et al.* (2021) ‘Needs, gaps and opportunities for standard and e-mental health care among at-risk populations in the Asia Pacific in the context of covid-19: A rapid scoping review’, *International Journal for Equity in Health*, 20(1). doi:10.1186/s12939-021-01484-5.

Murphy, K.P. (2012) *Machine learning: A probabilistic perspective*. Cambridge: The MIT Press.

Murray, E. *et al.* (2016) ‘Evaluating Digital Health Interventions’, *American Journal of Preventive Medicine*, 51(5), pp. 843–851. doi:10.1016/j.amepre.2016.06.008.

Nardelli, M. *et al.* (2015) “Recognizing emotions induced by affective sounds through heart rate variability,” *IEEE Transactions on Affective Computing*, 6(4), pp. 385–394. Available at: https://doi.org/10.1109/taffc.2015.2432810.

Neves, I. *et al.* (2021) ‘Interpretable heartbeat classification using local model-agnostic explanations on ecgs’, *Computers in Biology and Medicine*, 133, p. 104393. doi:10.1016/j.compbiomed.2021.104393.

Ng, A.Y. (2004) ‘Feature selection, l1 vs. l2 regularization, and rotational invariance’, *Twenty-first international conference on Machine learning - ICML ’04*, p. 78. doi:10.1145/1015330.1015435.

Ng, H.-W. *et al.* (2015) “Deep learning for emotion recognition on small datasets using transfer learning,” *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* [Preprint]. Available at: https://doi.org/10.1145/2818346.2830593.

Nonis, F. *et al.* (2020) ‘Questionnaires or Inner feelings: Who measures the engagement better?’, *Applied Sciences*, 10(2), p. 609. doi:10.3390/app10020609.

North, M.M., North, S.M. and Coble, J.R. (1998) ‘Virtual reality therapy: An effective treatment for the fear of public speaking’, *International Journal of Virtual Reality*, 3(3), pp. 1–6. doi:10.20870/ijvr.1998.3.3.2625.

Nuske, H.J. *et al.* (2014) “Pupillometry reveals reduced unconscious emotional reactivity in autism,” *Biological Psychology*, 101, pp. 24–35. Available at: https://doi.org/10.1016/j.biopsycho.2014.07.003.

O’Brien, H.L. and Toms, E.G. (2008) ‘What is user engagement? A conceptual framework for defining user engagement with technology’, *Journal of the American Society for Information Science and Technology*, 59(6), pp. 938–955. doi:10.1002/asi.20801.

Öst, L.-G., Brandberg, M. and Alm, T. (1997) ‘One versus five sessions of exposure in the treatment of flying phobia’, *Behaviour Research and Therapy*, 35(11), pp. 987996. doi:10.1016/s0005-7967(97)00077-6.

Pal, S.K. and Mitra, S. (1992) ‘Multilayer Perceptron, fuzzy sets, and classification’, *IEEE Transactions on Neural Networks*, 3(5), pp. 683–697. doi:10.1109/72.159058.

Palatnik de Sousa, I., Maria Bernardes Rebuzzi Vellasco, M. and Costa da Silva, E. (2019) ‘Local interpretable model-agnostic explanations for classification of lymph node metastases’, *Sensors*, 19(13), p. 2969. doi:10.3390/s19132969.

Panda, R., Malheiro, R. and Paiva, R.P. (2020) “Novel audio features for Music Emotion Recognition,” *IEEE Transactions on Affective Computing*, 11(4), pp. 614–626. Available at: https://doi.org/10.1109/taffc.2018.2820691.

Park, C.Y. *et al.* (2020) “K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations,” *Scientific Data*, 7(1). Available at: https://doi.org/10.1038/s41597-020-00630-y.

Park, K.-M. *et al.* (2011) ‘A virtual reality application in role-plays of social skills training for schizophrenia: A randomized, controlled trial’, *Psychiatry Research*, 189(2), pp. 166–172. doi:10.1016/j.psychres.2011.04.003.

Parsons, T.D. and Rizzo, A.A. (2008) ‘Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: A meta-analysis’, *Journal of Behavior Therapy and Experimental Psychiatry*, 39(3), pp. 250–261. doi:10.1016/j.jbtep.2007.07.007.

Pawar, U., O'Shea, D., Rea, S., O'Reilly, R., 2020. Incorporating Explainable Artificial Intelligence (XAI) to aid the Understanding of Machine Learning in the Healthcare Domain.

Pedregosa, F. et al. (2011) ‘Scikit-learn: Machine Learning in Python’, Journal of Machine Learning Research, 12, pp. 2825–2830.

Pedrelli, P. *et al.* (2020) 'Monitoring changes in depression severity using wearable and mobile sensors', *Frontiers in Psychiatry*, 11. doi:10.3389/fpsyt.2020.584711.

Persky, S. (2011) ‘Application of virtual reality methods to Obesity Prevention and Management Research’, *Journal of Diabetes Science and Technology*, 5(2), pp. 333–339. doi:10.1177/193229681100500220.

Petmezas, G. *et al.* (2021) ‘Automated atrial fibrillation detection using a hybrid CNN-LSTM network on imbalanced ECG datasets’, *Biomedical Signal Processing and Control*, 63, p. 102194. doi:10.1016/j.bspc.2020.102194.

Pham, T.D. (2021) “Time–Frequency Time–space LSTM for robust classification of physiological signals,” *Scientific Reports*, 11(1). Available at: https://doi.org/10.1038/s41598-021-86432-7.

Philippe, T.J. *et al.* (2022) ‘Digital Health Interventions for delivery of mental health care: Systematic and comprehensive meta-review’, *JMIR Mental Health*, 9(5).doi:10.2196/35159.

PHQ-9 Depression Test Questionnaire (2021) *Patient.info*. Available at: https://patient.info/doctor/patient-health-questionnaire-phq-9 (Accessed: March 10, 2021).

Picard, R.W. (2009) “Future affective technology for autism and Emotion Communication,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535), pp. 3575–3584. Available at: https://doi.org/10.1098/rstb.2009.0143.

Plasqui, G., Bonomi, A.G. and Westerterp, K.R. (2013) ‘Daily physical activity assessment with accelerometers: New insights and validation studies’, *Obesity Reviews*, 14(6), pp. 451–462. doi:10.1111/obr.12021.

Poplin, R., Varadarajan, A., Blumer, K., Liu, Y., McConnell, M., Corrado, G., Peng, L. and Webster, D., 2018. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3), pp.158-164.

POSNER, J., RUSSELL, J.A. and PETERSON, B.S. (2005) ‘The Circumplex model of affect: An integrative approach to Affective Neuroscience, Cognitive Development, and psychopathology’, *Development and Psychopathology*, 17(03). doi:10.1017/s0954579405050340.

Pot-Kolder, R.M. *et al.* (2018) ‘Virtual-reality-based cognitive behavioural therapy versus waiting list control for paranoid ideation and social avoidance in patients with psychotic disorders: A single-blind randomised controlled trial’, *The Lancet Psychiatry*, 5(3), pp. 217–226. doi:10.1016/s2215-0366(18)30053-1.

Preece, A., Harborne, D., Braines, D., Tomsett, R. and Chakraborty, S., 2018. Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184*.

Probst, P., Wright, M.N. and Boulesteix, A.L. (2019) “Hyperparameters and tuning strategies for Random Forest,” *WIREs Data Mining and Knowledge Discovery*, 9(3). Available at: https://doi.org/10.1002/widm.1301.

ProPublica. 2020. *Bias In Criminal Risk Scores Is Mathematically Inevitable, Researchers Say*. [online] Available at: <https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say> [Accessed 19 June 2020].

Psaltis, A. *et al.* (2018) ‘Multimodal student engagement recognition in Prosocial Games’, *IEEE Transactions on Games*, 10(3), pp. 292–303. doi:10.1109/tciaig.2017.2743341.

Pu Liang et al., “MultiViz: Towards visualizing and understanding multimodal models,” 2022, *arXiv:2207.00056*.

Qu, Q.-X., Guo, F. and Duffy, V.G. (2017) “Effective use of human physiological metrics to evaluate website usability,” *Aslib Journal of Information Management*, 69(4), pp. 370–388. Available at: https://doi.org/10.1108/ajim-09-2016-0155.

Rahman, M.A. *et al.* (2022) ‘Explainable multimodal machine learning for engagement analysis by continuous performance test’, *Universal Access in Human-Computer Interaction. User and Context Diversity*, pp. 386–399. doi:10.1007/978-3-031-05039-8_28.

Rai, H.M., Chatterjee, K. and Mukherjee, C. (2020) ‘Hybrid CNN-LSTM model for automatic prediction of cardiac arrhythmias from ECG Big Data’, *2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)* [Preprint]. doi:10.1109/upcon50219.2020.9376450.

Rani, P. *et al.* (2006) ‘An empirical study of machine learning techniques for affect recognition in human–robot interaction’, *Pattern Analysis and Applications*, 9(1), pp. 58–69. doi:10.1007/s10044-006-0025-y.

Raudonis, V. *et al.* (2013) ‘Evaluation of human emotion from Eye Motions’, *International Journal of Advanced Computer Science and Applications*, 4(8). doi:10.14569/ijacsa.2013.040812.

Reeves, B., & Nass, C. I. 1996. *The media equation: How people treat computers, television, and new media like real people and places.* Center for the Study of Language and Information; Cambridge University Press.

Reger, G.M. *et al.* (2011) “Effectiveness of virtual reality exposure therapy for active duty soldiers in a military mental health clinic,” *Journal of Traumatic Stress*, 24(1), pp. 93–96. Available at: https://doi.org/10.1002/jts.20574.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) ‘“why should I trust you?”’, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* [Preprint]. doi:10.1145/2939672.2939778.

Ricci, S. *et al.* (2022) ‘Viewpoint: Virtual and augmented reality in basic and Advanced Life Support training’, *JMIR Serious Games*, 10(1). doi:10.2196/28595.

Rim, B. *et al.* (2020) ‘Deep learning in Physiological Signal Data: A survey’, *Sensors*, 20(4), p. 969. doi:10.3390/s20040969.

Riva, G. *et al.* (2007) ‘Affective interactions using virtual reality: The link between presence and emotions’, *CyberPsychology &amp; Behavior*, 10(1), pp. 45–56. doi:10.1089/cpb.2006.9993.

Rizzo, A. *et al.* (2009) “Virtual reality exposure therapy for combat-related PTSD,” *Post-Traumatic Stress Disorder*, pp. 375–399. Available at: https://doi.org/10.1007/978-1-60327-329-9_18.

Roberts, L.W., Chan, S. and Torous, J. (2018) ‘New tests, new tools: Mobile and connected technologies in advancing psychiatric diagnosis’, *npj Digital Medicine*, 1(1). doi:10.1038/s41746-017-0006-0.

Robinson, N. (2018) *10 minutes of Tetris Effect Music and gameplay*, *YouTube*. Available at: https://www.youtube.com/watch?v=urbLIyd-VsQ (Accessed: 15 September 2023).

Rodríguez-Hernández, M. *et al.* (2021) ‘Effects of specific virtual reality-based therapy for the rehabilitation of the upper limb motor function post-ictus: Randomized controlled trial’, *Brain Sciences*, 11(5), p. 555. doi:10.3390/brainsci11050555.

Rogers, J.M. *et al.* (2020) ‘Single-channel EEG measurement of engagement in virtual rehabilitation: A validation study’, *Virtual Reality*, 25(2), pp. 357–366. doi:10.1007/s10055-020-00460-8.

Roh, T., Sunjoo Hong and Hoi-Jun Yoo, 2014. Wearable depression monitoring system with heart-rate variability. *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*,.

Rothbaum, B.O. *et al.* (1999) ‘Virtual reality exposure therapy for PTSD Vietnam veterans: A case study’, *Journal of Traumatic Stress*, 12(2), pp. 263–271. doi:10.1023/a:1024772308758.

Roumia, M. and Steinhubl, S. (2014) ‘Improving cardiovascular outcomes using Electronic Health Records’, *Current Cardiology Reports*, 16(2). doi:10.1007/s11886-013-0451-6.

Rude, S., Gortner, E. and Pennebaker, J., 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8), pp.1121-1133.

Ruiz-del-Solar, J. *et al.* (2021)' Mental and emotional health care for covid-19 patients: Employing pudu, a telepresence robot', *IEEE Robotics &amp; Automation Magazine*, 28(1), pp. 82–89. doi:10.1109/mra.2020.3044906.

Ruqeyya, G. *et al.* (2022) 'EEG-based engagement index for Video Game Players', *2022 International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE)* [Preprint]. doi:10.1109/etecte55893.2022.10007386.

Rus-Calafell, M. *et al.* (2017) 'Virtual reality in the assessment and treatment of psychosis: A systematic review of its utility, acceptability and effectiveness', *Psychological Medicine*, 48(3), pp. 362–391. doi:10.1017/s0033291717001945.

Russell, J.A. (1980) "A circumplex model of affect.," *Journal of Personality and Social Psychology*, 39(6), pp. 1161–1178. Available at: https://doi.org/10.1037/h0077714.

Saeb, S. *et al.* (2015) 'Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study', *Journal of Medical Internet Research*, 17(7). doi:10.2196/jmir.4273.

Said, N.S. (2004) 'An engaging multimedia design model', *Proceedings of the 2004 conference on Interaction design and children: building a community* [Preprint]. doi:10.1145/1017833.1017873.

Salvucci, D.D. and Goldberg, J.H. (2000) "Identifying fixations and saccades in eye-tracking protocols," *Proceedings of the symposium on Eye tracking research & applications - ETRA '00* [Preprint]. Available at: https://doi.org/10.1145/355017.355028.

Samoilov, A. and Goldfried, M.R. (2000) 'Role of emotion in cognitive-behavior therapy.', *Clinical Psychology: Science and Practice*, 7(4), pp. 373–385. doi:10.1093/clipsy.7.4.373.

Sandberg, C.E. *et al.* (2019) 'Using telemedicine to diagnose surgical site infections in low- and middle-income countries: Systematic Review', *JMIR mHealth and uHealth*, 7(8). doi:10.2196/13309.

Santamaria-Granados, L. *et al.* (2019) "Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos)," *IEEE Access*, 7, pp. 57–67. Available at: https://doi.org/10.1109/access.2018.2883213.

Sarkar, P. and Etemad, A. (2022) "Self-supervised ECG Representation Learning for emotion recognition," *IEEE Transactions on Affective Computing*, 13(3), pp. 1541–1554. Available at: https://doi.org/10.1109/taffc.2020.3014842.

Sato, W., Kochiyama, T. and Yoshikawa, S. (2020) ‘Physiological correlates of subjective emotional valence and arousal dynamics while viewing films’, *Biological Psychology*, 157, p. 107974. doi:10.1016/j.biopsycho.2020.107974.

Sau, A. and Bhakta, I. (2017) 'Predicting anxiety and depression in elderly patients using machine learning technology', *Healthcare Technology Letters*, 4(6), pp. 238–243. doi:10.1049/htl.2016.0096.

Sayani, S. *et al.* (2019) ‘Addressing cost and time barriers in chronic disease management through telemedicine: An exploratory research in select low- and middle-income countries’, *Therapeutic Advances in Chronic Disease*, 10, p. 204062231989158. doi:10.1177/2040622319891587.

Schiratti, J.-B. *et al.* (2018) ‘An ensemble learning approach to detect epileptic seizures from long intracranial EEG Recordings’, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* [Preprint]. doi:10.1109/icassp.2018.8461489.

Schmidt, P. *et al.* (2018) “Introducing WESAD, a multimodal dataset for wearable stress and affect detection,” *Proceedings of the 20th ACM International Conference on Multimodal Interaction* [Preprint]. Available at: https://doi.org/10.1145/3242969.3242985.

Schuch, F. *et al.* (2017) "Physical activity and sedentary behavior in people with major depressive disorder: A systematic review and meta-analysis," *Journal of Affective Disorders*, 210, pp. 139–150. Available at: https://doi.org/10.1016/j.jad.2016.10.050.

Segawa, T. *et al.* (2020) ‘Virtual reality (VR) in assessment and treatment of addictive disorders: A systematic review’, *Frontiers in Neuroscience*, 13. doi:10.3389/fnins.2019.01409.

Segler, M.H. *et al.* (2017) ‘Generating focused molecule libraries for drug discovery with recurrent neural networks’, *ACS Central Science*, 4(1), pp. 120–131. doi:10.1021/acscentsci.7b00512.

Selvaraju, R.R. *et al.* (2017) ‘Grad-cam: Visual explanations from deep networks via gradient-based localization’, *2017 IEEE International Conference on Computer Vision (ICCV)* [Preprint]. doi:10.1109/iccv.2017.74.

Senbekov, M. *et al.* (2020) 'The recent progress and applications of digital technologies in Healthcare: A Review', *International Journal of Telemedicine and Applications*, 2020, pp. 1–18. doi:10.1155/2020/8830200.

Sequeira, H. *et al.* (2009) "Electrical autonomic correlates of emotion," *International Journal of Psychophysiology*, 71(1), pp. 50–56. Available at: https://doi.org/10.1016/j.ijpsycho.2008.07.009.

Setianto, E.J. *et al.* (2022) 'Eye tracking and emotion recognition using multiple spatial-temporal networks', *2022 International Conference on Data Science and Its Applications (ICoDSA)* [Preprint]. doi:10.1109/icodsa55874.2022.9862881.

Sharma, N., Shamkuwar, M. and Singh, I. (2018) 'The history, present and future with IOT', *Intelligent Systems Reference Library*, pp. 27–51. doi:10.1007/978-3-030-04203-5_3.

Sharma, R. and Kshetri, N. (2020) 'Digital Healthcare: Historical Development, applications, and future research directions', *International Journal of Information Management*, 53, p. 102105. doi:10.1016/j.ijinfomgt.2020.102105.

Sherstinsky, A. (2020) 'Fundamentals of Recurrent Neural Network (RNN) and long short-term memory (LSTM) network', *Physica D: Nonlinear Phenomena*, 404, p. 132306. doi:10.1016/j.physd.2019.132306.

Shickel, B., Tighe, P., Bihorac, A. and Rashidi, P., 2018. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5), pp.1589-1604.

Shudes, C. *et al.* (2023) *Digital Transformation*, *Deloitte Insights*. Available at: https://www2.deloitte.com/us/en/insights/industry/health-care/digital-transformation-in-healthcare.html (Accessed: 14 December 2023).

Shukla, J. *et al.* (2021) "Feature extraction and selection for emotion recognition from electrodermal activity," *IEEE Transactions on Affective Computing*, 12(4), pp. 857–869. Available at: https://doi.org/10.1109/taffc.2019.2901673.

Siami-Namini, S., Tavakoli, N. and Siami Namin, A. (2018) ‘A comparison of Arima and LSTM in forecasting time series’, *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* [Preprint]. doi:10.1109/icmla.2018.00227.

Siegel, A. *et al.* (2021) ‘Barriers, benefits and interventions for improving the delivery of telemental health services during the coronavirus disease 2019 pandemic: A systematic review’, *Current Opinion in Psychiatry*, 34(4), pp. 434–443. doi:10.1097/yco.0000000000000714.

Sivertsen, B., Krokstad, S., Øverland, S. and Mykletun, A., 2009. The epidemiology of insomnia: Associations with physical and mental health. *Journal of Psychosomatic Research*, 67(2), pp.109-116.

Sloan, D.M. and Kring, A.M. (2007) 'Measuring changes in emotion during psychotherapy: Conceptual and methodological issues.', *Clinical Psychology: Science and Practice*, 14(4), pp. 307–322. doi:10.1111/j.1468-2850.2007.00092.x.

Sloan, D.M. and Kring, A.M. (2007) ‘Measuring changes in emotion during psychotherapy: Conceptual and methodological issues.’, *Clinical Psychology: Science and Practice*, 14(4), pp. 307–322. doi:10.1111/j.1468-2850.2007.00092.x.

Snippe, E. *et al.* (2018) ‘Explaining variability in therapist adherence and patient depressive symptom improvement: The role of therapist interpersonal skills and patient engagement’, *Clinical Psychology &amp; Psychotherapy*, 26(1), pp. 84–93. doi:10.1002/cpp.2332.

Sogo, H. (2012) “GazeParser: An open-source and multiplatform library for low-cost eye tracking and analysis,” *Behavior Research Methods*, 45(3), pp. 684–695. Available at: https://doi.org/10.3758/s13428-012-0286-x.

Soleymani, M. *et al.* (2012) “A multimodal database for affect recognition and implicit tagging,” *IEEE Transactions on Affective Computing*, 3(1), pp. 42–55. Available at: https://doi.org/10.1109/t-affc.2011.25.

Soleymani, M. *et al.* (2014) ‘Continuous emotion detection using EEG signals and facial expressions’, *2014 IEEE International Conference on Multimedia and Expo (ICME)* [Preprint]. doi:10.1109/icme.2014.6890301.

Statista Research Department, statista and 1, S. (2023) *Ownership of smartphones in the UK 2022*, *Statista*. Available at: https://www.statista.com/statistics/956297/ownership-of-smartphones-uk/ (Accessed: 04 January 2024).

Statista. 2020. *UK: Social Media Usage 2019 | Statista*. [online] Available at: <https://www.statista.com/statistics/507405/uk-active-social-media-and-mobile-social-media-users/> [Accessed 7 July 2020].

Štrumbelj, E. and Kononenko, I., 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, *41*, pp.647-665.

Suh, A. and Prophet, J. (2018) 'The state of immersive technology research: A literature analysis', *Computers in Human Behavior*, 86, pp. 77–90. doi:10.1016/j.chb.2018.04.019.

Sümer, Ö. *et al.* (2023) 'Multimodal engagement analysis from facial videos in the classroom', *IEEE Transactions on Affective Computing*, 14(2), pp. 1012–1027. doi:10.1109/taffc.2021.3127692.

Tabbaa, L. *et al.* (2021) "Vreed," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(4), pp. 1–20. Available at: https://doi.org/10.1145/3495002.

Tang, H. *et al.* (2017) "Multimodal emotion recognition using Deep Neural Networks," *Neural Information Processing*, pp. 811–819. Available at: https://doi.org/10.1007/978-3-319-70093-9_86.

Tao, L.-Y. and Lu, B.-L. (2020) "Emotion recognition under sleep deprivation using a multimodal residual LSTM network," *2020 International Joint Conference on Neural Networks (IJCNN)* [Preprint]. Available at: https://doi.org/10.1109/ijcnn48605.2020.9206957.

Tetris® effect: Connected (2023) *TETRIS® EFFECT: CONNECTED*. Available at: https://www.tetriseffect.game/ (Accessed: 15 September 2023).

Teychenne, M., Ball, K. and Salmon, J. (2008) "Physical activity and likelihood of depression in adults: A Review," *Preventive Medicine*, 46(5), pp. 397–411. Available at: https://doi.org/10.1016/j.ypmed.2008.01.009.

Theissler, A. *et al.* (2022) 'Explainable AI for Time Series classification: A review, Taxonomy and Research Directions', *IEEE Access*, 10, pp. 100700–100724. doi:10.1109/access.2022.3207765.

Tjoa, E. and Guan, C. (2021) ‘A survey on Explainable Artificial Intelligence (XAI): Toward medical xai’, *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), pp. 4793–4813. doi:10.1109/tnnls.2020.3027314.

Tong, H.L. and Laranjo, L. (2018) ‘The use of social features in mobile health interventions to promote physical activity: A systematic review’, *npj Digital Medicine*, 1(1). doi:10.1038/s41746-018-0051-3.

Trappey, A. *et al.* (2020) ‘Virtual reality exposure therapy for driving phobia disorder (2): System refinement and verification’, *Applied Sciences*, 11(1), p. 347. doi:10.3390/app11010347.

Tripathi, S. *et al.* (2017) “Using deep and convolutional neural networks for accurate emotion classification on DEAP Data,” *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(2), pp. 4746–4752. Available at: https://doi.org/10.1609/aaai.v31i2.19105.

Tropsha, A., 2010. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics*, 29(6-7), pp.476-488.

Truss, C. (2014) *Employee engagement in theory and Practice*. Milton Park, Abingdon, Oxon: Routledge.

Tsang, V. (2016) ‘Eye-tracking study on facial emotion recognition tasks in individuals with high-functioning autism spectrum disorders’, *Autism*, 22(2), pp. 161–170. doi:10.1177/1362361316667830.

Tseng, P.-Y. *et al.* (2020) ‘Prediction of the development of acute kidney injury following cardiac surgery by Machine Learning’, *Critical Care*, 24(1). doi:10.1186/s13054-020-03179-9.

TUDOR-LOCKE, CATRINE, JOHNSON, W.I.L.L.I.A.M.D. and KATZMARZYK, PETERT (2009) "Accelerometer-determined steps per day in adults," *Medicine & Science in Sports & Exercise*, 41(7), pp. 1384–1391. Available at: https://doi.org/10.1249/mss.0b013e318199885c.

Tutun, S. *et al.* (2022) ‘An AI-based decision support system for predicting mental health disorders’, *Information Systems Frontiers*, 25(3), pp. 1261–1276. doi:10.1007/s10796-022-10282-5.

Udovičić, G. *et al.* (2017) ‘Wearable emotion recognition system based on GSR and PPG signals’, *Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care* [Preprint]. doi:10.1145/3132635.3132641.

Udovičić, G. *et al.* (2017) “Wearable emotion recognition system based on GSR and PPG signals,” *Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care* [Preprint]. Available at: https://doi.org/10.1145/3132635.3132641.

Udrea, A., Mitra, G., Costea, D., Noels, E., Wakkee, M., Siegel, D., Carvalho, T. and Nijsten, T., 2019. Accuracy of a smartphone application for triage of skin lesions based on machine learning algorithms. *Journal of the European Academy of Dermatology and Venereology*, 34(3), pp.648-655.

Usmani, A., Imran, M. and Javaid, Q. (2022) ‘Usage of artificial intelligence and virtual reality in medical studies’, *Pakistan Journal of Medical Sciences*, 38(4). doi:10.12669/pjms.38.4.5910.

Valenza, G., Lanata, A. and Scilingo, E.P. (2012) “The Role of Nonlinear Dynamics in affective valence and arousal recognition,” *IEEE Transactions on Affective Computing*, 3(2), pp. 237–249. Available at: https://doi.org/10.1109/t-affc.2011.30.

Vallance, J., Winkler, E., Gardiner, P., Healy, G., Lynch, B. and Owen, N., 2011. Associations of objectively-assessed physical activity and sedentary time with depression: NHANES (2005–2006). *Preventive Medicine*, 53(4-5), pp.284-288.

Van Gent, P. *et al.* (2019) “HeartPy: A novel heart rate algorithm for the analysis of noisy signals,” *Transportation Research Part F: Traffic Psychology and Behaviour*, 66, pp. 368–378. Available at: https://doi.org/10.1016/j.trf.2019.09.015.

Ventura, R.B. and Porfiri, M. (2020) ‘Galvanic skin response as a measure of engagement during play in virtual reality’, *Volume 1: Adaptive/Intelligent Sys. Control; Driver Assistance/Autonomous Tech.; Control Design Methods; Nonlinear Control; Robotics; Assistive/Rehabilitation Devices; Biomedical/Neural Systems; Building Energy Systems; Connected Vehicle Systems; Control/Estimation of Energy Systems; Control Apps.; Smart Buildings/Microgrids; Education; Human-Robot Systems; Soft Mechatronics/Robotic Components/Systems; Energy/Power Systems; Energy Storage; Estimation/Identification; Vehicle Efficiency/Emissions* [Preprint]. doi:10.1115/dscc2020-3177.

Vielhaben, J. *et al.* (2024) ‘Explainable AI for time series via Virtual Inspection Layers’, *Pattern Recognition*, 150, p. 110309. doi:10.1016/j.patcog.2024.110309.

Wahle, F. *et al.* (2016) "Mobile Sensing and support for people with depression: A pilot trial in the wild," *JMIR mHealth and uHealth*, 4(3). Available at: https://doi.org/10.2196/mhealth.5960.

Wang, J. and Wang, M. (2021) ‘Review of the emotional feature extraction and classification using EEG signals’, *Cognitive Robotics*, 1, pp. 29–40. doi:10.1016/j.cogr.2021.04.001.

Wang, R. *et al.* (2018) 'Tracking depression dynamics in college students using mobile phone and wearable sensing', *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1), pp. 1–26. doi:10.1145/3191775.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D. and Campbell, A., (2014). StudentLife. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '14 Adjunct*,.

Wang, X. *et al.* (2013) "A depression detection model based on sentiment analysis in micro-blog social network," *Lecture Notes in Computer Science*, pp. 201–213. Available at: https://doi.org/10.1007/978-3-642-40319-4_18.

Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L. and Bao, Z., 2013. A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network. *Lecture Notes in Computer Science*, pp.201-213.

Warmerdam, L. *et al.* (2010) ‘Cost-utility and cost-effectiveness of internet-based treatment for adults with depressive symptoms: Randomized trial’, *Journal of Medical Internet Research*, 12(5). doi:10.2196/jmir.1436.

Watson, N.F. *et al.* (2015) "Recommended amount of sleep for a healthy adult: A joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society," *Journal of Clinical Sleep Medicine*, 11(06), pp. 591–592. Available at: https://doi.org/10.5664/jcsm.4758.

Webster, P. (2020) ‘Virtual health care in the era of covid-19’, *The Lancet*, 395(10231), pp. 1180–1181. doi:10.1016/s0140-6736(20)30818-7.

Weir, K. (2021) *The age of digital interventions*, *Monitor on Psychology*. Available at: https://www.apa.org/monitor/2021/10/news-digital-interventions (Accessed: 22 December 2023).

Weiskopf, N., Hripcsak, G., Swaminathan, S. and Weng, C., 2013. Defining and measuring completeness of electronic health records for secondary use. *Journal of Biomedical Informatics*, 46(5), pp.830-836.

Welsh Health Survey 2015: Health status, illness, and other conditions.

What is exposure therapy? (2021) *American Psychological Association*. American Psychological Association. Available at: https://www.apa.org/ptsd-guideline/patients-and-families/exposure-therapy (Accessed: February 10, 2021).

Wiederhold, M.D. and Wiederhold, B.K. (2007) 'Virtual reality and interactive simulation for pain distraction', *Pain Medicine*, 8(suppl 3). doi:10.1111/j.1526-4637.2007.00381.x.

Wijsman, J., Grundlehner, B., Hao Liu, Hermens, H. and Penders, J., 2011. Towards mental stress detection using wearable physiological sensors. *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*,.

Wilson, E.B. (1927) 'Probable inference, the law of succession, and statistical inference', *Journal of the American Statistical Association*, 22(158), pp. 209–212. doi:10.1080/01621459.1927.10502953.

Witteveen, A.B. *et al.* (2022) 'Remote mental health care interventions during the COVID-19 pandemic: An Umbrella Review', *Behaviour Research and Therapy*, 159, p. 104226. doi:10.1016/j.brat.2022.104226.

Wundt, W.,Principles of physiological psychology, 1873,in:Readings in the history of psychology,EastNorwalk, CT, US: Appleton-Century-Crofts., 1948 pp. 248–250.

Yang, L. and Shami, A. (2020) 'On hyperparameter optimization of Machine Learning Algorithms: Theory and practice', *Neurocomputing*, 415, pp. 295–316. doi:10.1016/j.neucom.2020.07.061.

Yannakakis, G. 2018. Enhancing health care via affective computing. Malta Journal of Health Sciences 5, 1 (2018), 38–42. https://doi.org/10.14614/HEALTHCOMP/9/18

Yildirim, C. and O’Grady, T. (2020) ‘The efficacy of a virtual reality-based mindfulness intervention’, *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)* [Preprint]. doi:10.1109/aivr50618.2020.00035.

Yin, Y. *et al.* (2018) “Classification of eye tracking data using a convolutional neural network,” *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* [Preprint]. Available at: https://doi.org/10.1109/icmla.2018.00085.

Yuen, B., Dong, X. and Lu, T. (2019) ‘Inter-patient CNN-LSTM for QRS complex detection in noisy ECG signals’, *IEEE Access*, 7, pp. 169359–169370. doi:10.1109/access.2019.2955738.

Zhai, L., Zhang, H. and Zhang, D. (2015) "Sleep duration and depression among adults: A meta-analysis of prospective studies," *Depression and Anxiety*, 32(9), pp. 664–670. Available at: https://doi.org/10.1002/da.22386.

Zhang, C. *et al.* (2023) ‘Large language models for human–robot interaction: A Review’, *Biomimetic Intelligence and Robotics*, 3(4), p. 100131. doi:10.1016/j.birob.2023.100131.

Zhang, L., Tan, J., Han, D. and Zhu, H., 2017. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today*, 22(11), pp.1680-1685.

Zhang, S., Liu, G. and Lai, X. (2015) “Classification of evoked emotions using an artificial neural network based on single, short-term physiological signals,” *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 19(1), pp. 118–126. Available at: https://doi.org/10.20965/jaciii.2015.p0118.

Zhang, X. *et al.* (2020) ‘Artificial Intelligence Medical Ultrasound equipment: Application of breast lesions detection’, *Ultrasonic Imaging*, 42(4–5), pp. 191–202. doi:10.1177/0161734620928453.

Zhang, Z. *et al.* (2019) ‘Pathologist-level interpretable whole-slide cancer diagnosis with Deep Learning’, *Nature Machine Intelligence*, 1(5), pp. 236–245. doi:10.1038/s42256-019-0052-1.

Zhang, Z., Cui, L., Liu, X. and Zhu, T., (2016). Emotion Detection Using Kinect 3D Facial Points. *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*,.

Zhao, S., Zhao, Q., Zhang, X., Peng, H., Yao, Z., Shen, J., Yao, Y., Jiang, H. and Hu, B., 2017. Wearable EEG-Based Real-Time System for Depression Monitoring. *Brain Informatics*, pp.190-201.

Zheng, W.-L. *et al.* (2019) "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Transactions on Cybernetics*, 49(3), pp. 1110–1122. Available at: https://doi.org/10.1109/tcyb.2018.2797176.

Zhou, H. *et al.* (2021) 'Virtual reality as a reflection technique for public speaking training', *Applied Sciences*, 11(9), p. 3988. doi:10.3390/app11093988.

Zhou, Y. *et al.* (2019) 'Predictive big data analytics using the UK Biobank Data', *Scientific Reports*, 9(1). doi:10.1038/s41598-019-41634-y.

Zitouni, M.S. *et al.* (2023) "LSTM-modeling of emotion recognition using peripheral physiological signals in naturalistic conversations," *IEEE Journal of Biomedical and Health Informatics*, 27(2), pp. 912–923. Available at: https://doi.org/10.1109/jbhi.2022.3225330.

Zucco, C., Calabrese, B. and Cannataro, M. (2017) "Sentiment analysis and affective computing for Depression Monitoring," *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* [Preprint]. Available at: https://doi.org/10.1109/bibm.2017.8217966.

Zuroff, D.C. *et al.* (2016) 'Predictors and moderators of between-therapists and within-therapist differences in depressed outpatients' experiences of the Rogerian conditions.', *Journal of Counseling Psychology*, 63(2), pp. 162–172. doi:10.1037/cou0000139.

製品一覧: Fove official website (2021) *FOVE Official website | FOVE HealthCare*. Available at: https://fove-inc.com/product/ (Accessed: May 4, 2020).