CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Real-time anxiety classification in Virtual Reality Exposure Therapy
Authors	Mevlevioğlu, Deniz
Publication date	2025
Original Citation	Mevlevioğlu, D. 2025. Real-time anxiety classification in Virtual Reality Exposure Therapy. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2025, Deniz Mevlevioglu. - https://creativecommons.org/publicdomain/zero/1.0/
Download date	2026-11-07 00:23:50
Item downloaded from	https://hdl.handle.net/10468/17410

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Real-time Anxiety Classification in Virtual Reality Exposure Therapy

Deniz Mevlevioğlu
MSc
118224278



UNIVERSITY COLLEGE CORK

School of Computer Science and Information Technology

Thesis submitted for the degree of Doctor of Philosophy

February 2025

Head of Department: Prof Dirk Pesch

Supervisors: Dr Sabin Tabirca
Mr David Murphy

Research supported by Science Foundation Ireland

Contents

List of Figures . . . vi
List of Tables . . . viii
Abbreviations . . . ix
Acknowledgements . . . xiii
Abstract . . . xiv

1 Introduction **1**
1.1 Motivation . . . 1
1.1.1 Research Questions . . . 3
1.2 Contributions . . . 3
1.3 Publications . . . 4
1.4 Thesis Structure . . . 6

2 Scoping Literature Review **7**
2.1 Background . . . 7
2.1.1 Definition of Anxiety . . . 7
2.1.2 Anxiety Disorders . . . 8
2.1.2.1 Specific Phobias . . . 8
2.1.2.2 Social Anxiety Disorder . . . 9
2.1.2.3 Generalised Anxiety Disorder . . . 9
2.1.2.4 Panic Disorder . . . 10
2.1.3 Anxiety Response . . . 10
2.1.3.1 Psychophysiological Markers for Anxiety . . . 11
2.1.3.2 Behavioural Markers for Anxiety . . . 12
2.1.4 Sensor-based Anxiety Measurement . . . 13
2.1.5 Anxiety Classification using Machine Learning . . . 15
2.1.6 Non-VR Systems for Measuring Anxiety . . . 16
2.1.7 Exposure Therapy . . . 18
2.1.8 Virtual Reality Exposure Therapy . . . 19
2.1.8.1 Limitations of Virtual Reality . . . 21
2.1.8.2 Biosensors in Virtual Reality . . . 22
2.2 Scoping Review . . . 24
2.2.1 Application areas . . . 24
2.2.2 Research Questions . . . 25
2.3 Materials and Methods . . . 26
2.4 Results . . . 27
2.4.1 Physiological and Behavioural Measures . . . 28
2.4.2 Ground Truth . . . 31
2.4.3 Classification Models . . . 32
2.4.4 Gaps in the Literature . . . 34
2.5 Discussion . . . 34
2.5.1 Limitations . . . 35
2.5.2 Future Directions . . . 35
2.6 Conclusion . . . 36

3 eVRST: Emotional Virtual Reality Stroop Task Development 38
3.1 Background . . . 38
3.1.1 Anxiety classification and Ground Truth . . . 39
3.1.2 Stroop Colour-Word Task . . . 40
3.1.3 Emotional Stroop Colour-Word Task . . . 40
3.1.4 Commercial Devices used for Measuring Anxiety . . . 40
3.1.5 Related Work . . . 41
3.2 Design . . . 42
3.2.1 General Design . . . 42
3.2.2 3D Engine . . . 43
3.3 Implementation . . . 44
3.3.1 Stroop Scene . . . 44
3.3.1.1 Version History and Changes . . . 49
3.3.2 Nature Scene . . . 49
3.3.3 Audio . . . 50
3.3.3.1 Version History and Changes . . . 51
3.3.4 Interaction . . . 51
3.3.5 Tutorial . . . 52
3.3.5.1 Version History and Changes . . . 55
3.3.6 Accommodations . . . 55
3.3.7 Task . . . 55
3.3.7.1 Version History and Changes . . . 58
3.3.8 SUDS . . . 58
3.3.9 Biosignals . . . 59
3.3.9.1 Metrics and Devices . . . 60
3.3.9.2 Data Acquisition and Real-time Processing . . . 62
3.4 eVRST Testing . . . 63
3.5 Discussion . . . 63
3.5.1 Limitations . . . 65
3.6 Summary . . . 65

4 eVRST: Emotional Virtual Reality Stroop Task Validation 67
4.1 Background . . . 67
4.1.1 Hypothesis & Predictions . . . 68
4.1.2 Self-reported Anxiety . . . 68
4.1.3 Related Work . . . 69
4.2 Experiment Design & Methods . . . 70
4.2.1 Participants . . . 70
4.2.2 Materials . . . 71
4.2.3 Procedure . . . 71
4.2.4 Preparation of the data . . . 73
4.2.5 Performance-related Measures . . . 74
4.3 Descriptive Analysis . . . 75
4.3.1 Descriptive Statistics . . . 75
4.3.1.1 Recorded Variables . . . 75
4.3.1.2 Computed Variables . . . 76

4.4 Results . . . 80
4.4.1 Skin Conductivity . . . 82
4.4.2 Self-reported Anxiety . . . 82
4.4.3 Reaction Time . . . 82
4.4.4 Other Findings . . . 83
4.4.4.1 Effect of initial anxiety measured by STAI on performance and user state . . . 83
4.4.4.2 Relationship between different anxiety measures and reaction time . . . 84
4.4.4.3 ANOVA for demographic effects . . . 86
4.4.4.4 Comments . . . 87
4.5 Discussion . . . 87
4.5.1 Limitations . . . 89
4.5.2 Recommendations . . . 89
4.6 Conclusion . . . 90

5 Anxiety Classification **91**
5.1 Background . . . 91
5.1.1 Related Work . . . 94
5.2 Methods . . . 96
5.2.1 Dataset Creation . . . 96
5.2.2 Data Preparation . . . 97
5.2.2.1 Validation Splitting . . . 97
5.2.2.2 Normalisation . . . 97
5.2.2.3 Windowing . . . 97
5.2.3 Feature Extraction . . . 97
5.2.4 Feature Selection . . . 99
5.2.4.1 ANOVA . . . 99
5.2.4.2 Gini Index . . . 99
5.2.5 Feature Convergence . . . 100
5.3 Results . . . 102
5.3.1 Models . . . 102
5.3.1.1 Decision Trees . . . 102
5.3.1.2 Rule-based system . . . 103
5.3.1.3 K-Nearest Neighbour . . . 105
5.3.1.4 Support Vector Machine . . . 105
5.3.1.5 Artificial Neural Network . . . 105
5.3.1.6 Convolutional Neural Network . . . 106
5.3.2 Leave-One-Out Cross-Validation . . . 108
5.3.3 Window Selection . . . 109
5.3.4 Metric Comparison . . . 110
5.4 Discussion & Limitations . . . 111
5.4.1 Comparison to Similar Work . . . 114
5.4.2 Limitations . . . 115
5.4.3 Recommendations . . . 117
5.5 Conclusion . . . 117

6 Virtual Reality Exposure Therapy with Seamless Feedback: A Prototype **119**
6.1 Background . . . 120
6.1.1 Related Work . . . 121
6.2 Design . . . 122
6.2.1 Virtual Reality Exposure Environments . . . 122
6.2.1.1 Acrophobia . . . 122
6.2.1.2 Claustrophobia - Proof of concept . . . 125
6.2.1.3 Nature Scene . . . 125
6.2.2 Graphics . . . 127
6.2.3 Audio and Haptics . . . 128
6.3 Seamless Feedback Integration . . . 128
6.3.1 Unity solution . . . 128
6.3.2 Server solution . . . 130
6.3.2.1 Containerisation . . . 130
6.3.3 Comparison and Decision . . . 132
6.3.4 Real-Time Processing of Data . . . 132
6.4 UI . . . 133
6.4.1 Data Visualisation . . . 133
6.4.2 Level Manager . . . 134
6.5 Software Testing . . . 135
6.6 Summary . . . 135

7 Conclusions and Future Directions **137**
7.1 Summary . . . 137
7.2 Application areas . . . 139
7.3 Virtual Reality Mental Health Interventions Limitations . . . 139
7.4 Recommendations . . . 140
7.5 Final Words . . . 142

A PRISMA-ScR Checklist **169**

B Extraction Form - Scoping Review **172**

C Quality Assessment Scale **173**

D eStroop Code Snippets **176**

E Consent Form **181**

F Pre-Questionnaire **183**

G Post-Questionnaire **186**

H SPSS Variables **189**

I Formal SPSS Results **191**

J Framework Code Snippets **192**

K Affective Words **194**

List of Figures

2.1 Biosensors useful for measuring anxiety 14
2.2 The search results. WOS: Web of Science [PMB+21] 28
2.3 Quality scores of each study [MH20] 30

3.1 Emotional Stroop Task 41
3.2 eVRST task experience design 43
3.3 The waiting room scene, including the interaction platform . . . 45
3.4 Layout of the Waiting Room 46
3.5 Perspective of a user looking towards the waiting area 47
3.6 The reception desk 47
3.7 The VR slider used to log anxiety 48
3.8 The stages of the Stroop task 49
3.9 The nature scene 50
3.10 Task script options 53
3.11 eVRST task flow 56
3.12 Visual feedback for answers 59

4.1 Distribution of age groups among participants 71
4.2 Distribution of prior VR exposure among participants 71
4.3 A participant with the equipment attached 72
4.4 Distribution of STAI scores among participants 76
4.5 Distribution of log-transformed STAI scores 77
4.6 Distribution of square root transformed STAI scores 77
4.7 Distribution of SUDS increase between stages 1-3 79
4.8 Boxplot of SUDS increase 80
4.9 Boxplot of % SCL increase 80
4.10 Distribution of % SCL increase between stages 1-3 80
4.11 Boxplot of % SCL increase 81
4.12 Distribution of overall reaction time 81
4.13 Distribution of emotional Stroop effect 81
4.14 Mean SCL for baseline (0), stages 1, 2 and 3 83
4.15 ESE per each participant 84
4.16 Scatterplot between initial STAI scores and SCL increase 85
4.17 Scatterplot between SCL percentage increase and ESE 85
4.18 Boxplot of mean ESE by VR exposure 86

5.1 Left: total number of records with the keyword machine learning. Right: The proportion of machine learning records is compared to all records published in the same year. 92
5.2 Gini importances 101
5.3 Correlation Matrix between each feature 102
5.4 The ANN architecture 106
5.5 The CNN architecture 107

5.6 Classification accuracy for each participant using leave-one-out cross-validation ANN: Artificial Neural Network, CNN: Convolutional Neural Network. 109
5.7 Confusion matrix for the Convolutional Neural Network 0: No anxiety, 1: Mild anxiety, 2: Severe anxiety 110
5.8 Confusion matrices per signal combination 112

6.1 Framework Diagram. CNN: Convolutional Neural Network, EEG: Electroencephalogram GSR: Galvanic Skin Response, HMD: Head-mounted Display, PPG: Photoplethysmogram. 120
6.2 Perspective of the user looking out of the glass elevator, both eyes rendered 124
6.3 From top to bottom: Living room, big bedroom, small bedroom, box office and bathroom 126
6.4 Machine learning communication diagram 131
6.5 UI 134

List of Tables

2.1 Previous non-VR specific research on real-time anxiety prediction 17
2.2 Inclusion and exclusion criteria 26
2.3 Previous research on real-time anxiety prediction in VR 29

4.1 Previous work on Stroop and eStroop tasks 70
4.2 Descriptive statistics of SCL, RT and SUDS between each stage 78
4.3 Questionnaire responses 78
4.4 Descriptive statistics of computed information 79

5.1 The extracted features 98
5.2 The relationship between the feature and the condition 100
5.3 The relationship between the CNN features and the condition . 101
5.4 Baseline accuracies for each classifier 102
5.5 Time windows used and their corresponding classification accuracy 110
5.6 Comparison to similar studies 116

H.1 SPSS Variables 190

I.1 Mauchly's Test of Sphericity 191
I.2 ANOVA Univariate 191
I.3 Bonferroni pairwise comparisons of mean SCL between stages . 191

Abbreviations

AI	Artificial Intelligence
ACC	Accelerometer
ACM	Association for Computing Machinery
ANEW	Affective Norms for English Words
ANN	Artificial Neural Network
ANS	Autonomic Nervous System
ANOVA	Analysis of Variance
API	Application Programming Interface
BO	Blood Oximeter
BP	Blood Pressure
BPNN	Backpropagation Neural Network
BVP	Blood Volume Pressure
CAVE	Computer Automatic Virtual Environments
CBT	Cognitive-Behavioural Therapy
CDA	Continuous Decomposition Analysis
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DDA	Discrete Decomposition Analysis
DSS	Decision Support System
DT	Decision Tree
ECG	Electrocardiogram
EDA	Electrodermal Activity
EEG	Electroencephalogram
ELM	Extreme Learning Machine
EMDR	Eye Movement Desensitisation and Reprocessing
EMG	Electromyogram
ESE	Emotional Stroop Effect
eStroop	Emotional Stroop
eVRST	Emotional Virtual Reality Stroop Task
FIR	Finite Impulse Response
FOV	Field-of-view
FPS	Fear-potentiated Startle
GAD	Generalised Anxiety Disorder
GAD-7	Generalised Anxiety Disorder Scale
GBT	Gradient Boosted Tree
GSR	Galvanic Skin Response

GPU	Graphics Processing Unit
HPA	Hypothalamic Pituitary Adrenal
HR	Heart Rate
HRV	Heart Rate Variability
HMD	Head-Mounted Display
IAPS	the International Affective Picture System
IBI	Inter-beat Intervals
ID	Identification
IEEE	Institute of Electrical and Electronics Engineers
K-ELM	Kernel-Based Extreme Learning Machine
KNN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
LSL	LabStreamingLayer
LSTM	Long Short Term Memory
LTS	Long Term Support
LOOCV	Leave-One-Out Cross-Validation
MASQ	Mood and Anxiety Symptom Scale
ML	Machine Learning
MLP	Multi-Layer Perceptron
MPFC	Medial Prefrontal Cortex
NN	Normal-to-normal
PANAS	Positive and Negative Affect Schedule
PNS	Parasympathetic Nervous System
PPG	Photoplethysmogram
PR	Pulse Rate
PRV	Pulse Rate Variability
RCT	Randomised Controlled Trials
RDF	Random Decision Forest
RESP	Respiration
RT	Reaction Time
SD	Standard Deviation
SE	Stroop Effect
SAD	Social Anxiety Disorder
SAM	Sympathetic Adrenomedullary System
SCWT	Stroop Colour-Word Task
SKT	Skin Temperature
SNRI	Serotonin-Norepinephrine Reuptake Inhibitors

SNS	Sympathetic Nervous System
SSRI	Selective Serotonin Reuptake Inhibitors
STAI-Y1	State-Trait Anxiety Inventory - Y1
SUDS	Subjective Units of Distress Scale
SVM	Support Vector Machine
TP	Task Preparation
UI	User Interface
UST	Ultra-Short-Term
UX	User Experience
VE	Virtual Environment
VR	Virtual Reality
VRET	Virtual Reality Exposure Therapy
XGB	Extreme Gradient Boosting

I, Deniz Mevlevioğlu, certify that this thesis is my own work and I have not obtained a degree in this university or elsewhere on the basis of the work submitted in this thesis.

Deniz Mevlevioğlu

Acknowledgements

I would like to express my appreciation and gratitude to Sabin Tabirca and David Murphy for their supervision and support throughout my PhD research. I would also like to thank Dirk Pesch, Luis Gómez de Membrillera and James Duggan for their support through ADVANCE CRT. Additionally, I want to thank my colleagues in UCC and ADVANCE CRT for being an essential part of my PhD journey, especially Isabel Ronan and Ştefan-Bogdan Marcu.

I would also like to extend my gratitude to the participants who took part in my experiments. Lastly, I would like to send my appreciation and gratitude to my family and friends for their continuous support and encouragement.

This thesis has emanated from research supported in part by a Grant from Science Foundation Ireland under Grant number 18/CRT/6222.

Abstract

Virtual Reality Exposure Therapy (VRET) extends the reach of anxiety disorder treatment options by making them accessible and cost-efficient. Past research has supported the effectiveness of VRET in treating various disorders, such as generalised anxiety disorder and specific phobias. Based on the results of a scoping review, feedback options during therapy are currently inaccessible, and the limitations of biased self-report systems can hinder immersion into the virtual reality environment. Therefore, this thesis proposes a decision-support framework integrating continuous, seamless feedback on patient state into VRET. On-body physiological sensors are used to collect information on the patient's heart rate, skin conductivity and electrical brain activity. Ergonomic challenges arise while using such sensors with movement in Virtual Reality, and noise is introduced into signals from the electrical interference of the head-mounted display. Another challenge is establishing the ground truth for anxiety levels, which varies widely among past research. The proposed solution uses multi-modal signal acquisition combined with state of the art noise removal techniques to account for these limitations. The ground truth of anxiety is established by developing a Virtual Reality adaptation of the emotional Stroop task. The proposed emotional Virtual Reality Stroop Task (eVRST) is a cognitive task that requires users to log the colours of the words they see. High negative affect and high arousal words are chosen to evoke feelings of anxiety in users. The task has been split into three conditions to represent three distinct anxiety levels. A convolutional neural network is trained and deployed to classify the current anxiety state of the patient between no anxiety, mild anxiety and severe anxiety conditions with 75% accuracy. These calculations are carried out every five seconds and displayed to the therapist through a user interface that is not visible to the patient. Feedback from machine learning and the visualisation of raw signals help therapists make meaningful decisions and personalise the therapy to the patient's ever-changing needs. The feedback framework is transferrable, easily deployed, and scalable. In this thesis, recommendations for Virtual Reality bio-feedback systems are also highlighted to aid future research.

Chapter 1

Introduction

1.1 Motivation

Virtual reality (VR) has recently become increasingly popular due to its effectiveness, affordability, and accessibility. VR is a tool that has the potential to bring mental health care costs down and enhance the wellbeing of society. Thus, improving the state of the art VR therapy capabilities means securing more reliable therapy methods for those in need. The research in this field builds on the efficacy and effectiveness of such practices and aims to improve therapy outcomes for the public.

There are several caveats to the continuous use of VR therapeutically. First and foremost, VR use is not yet widespread enough to be intuitive. Factors such as age, education, and technical ability play an essential role in learning VR efficiently. Being unaccustomed to VR can decrease the value of VR therapy. Furthermore, some accessibility problems can make it more difficult for specific demographics, such as persons with visual or auditory impairments.

As VR is self-contained, it can also be quite challenging to communicate with the patient and receive feedback on the perceived effectiveness of therapy. Interruptions to the VR experience can be distracting, leading to diminished immersion in the treatment. A possible solution to this is the integration of self-reported feedback into the VR environment. However, doing so may introduce some bias due to the challenges introduced by the use of self-report measures.

This thesis showcases a system that seamlessly integrates objective feedback

into VR therapy environments. The solution involves two components. The first component strives to create a VR environment that can incorporate and use different therapy methods within the same system. Second, the system uses machine learning classification models to classify anxiety based on physiological and behavioural information. The next aim is to combine these systems to provide continuous real-time anxiety classification as a means of objective feedback while allowing mental health professionals to adjust many different methods of therapy based on these criteria using the framework.

The goals of this thesis were broken down into three major points. The first is measuring a reliable way of getting an accurate ground truth for anxiety. Determining the ground truth requires extensive research of past methods for determining the least biased way of measuring anxiety into different levels. This comprehensive research informs the design and implementation of a VR application of such techniques to be used effectively in VR therapy administration. This method's design, implementation and validation are crucial for a sound basis for the following steps and the real-time classification of anxiety.

The second point is the development of an algorithm to classify anxiety levels using biosensors and machine learning. This algorithm must be precise and accurate without hindering the performance of the VR environment or decreasing the quality of experience. A dataset must be created using the ground truth environment to develop such an algorithm. The time-series data needs to be normalised per participant, cleaned and filtered. The optimal features must be extracted to determine the best classification method.

The third point is the development of the therapy framework. A VR therapy environment must be built to administer the necessary stressing effect to be efficiently used as part of cognitive behavioural therapy. This therapy environment needs to be adjustable to suit the needs of different stages of the associated disorder, as well as the needs of other people. After that, a therapy framework must be created that will allow the addition of multiple types of therapy. This addition makes the therapy transferable for varying needs and use cases. These additional therapy environments must follow a predetermined set of rules, which must also be crafted. The predetermined set of rules would allow for a seamless transition between different types of therapy without making new applications every time or restarting the environment each time.

1.1.1 Research Questions

This thesis aims to answer the following research questions and subquestions:

- RQ1: How can uninterrupted and objective real-time feedback on patient anxiety be obtained during Virtual Reality Exposure Therapy?
- sub-RQ1: Can ground truth in anxiety be made more bias-free and accessible via the use of cognitive tasks?
- sub-RQ2: How accurately can anxiety be classified in real-time Virtual Reality using biosensors?
- sub-RQ3: Can anxiety classification be smoothly integrated with VR therapy environments to have a scalable and transferrable decision support framework?

1.2 Contributions

This work presents two main contributions to the literature on anxiety classification in Virtual Reality Therapy. The first is a method for detecting ground truth, named the emotional Stroop Colour-Word Task. It is the first Virtual Reality application of the long-established emotional Stroop Task, validated to work similarly in Virtual Reality compared to its previous in vivo applications. Furthermore, the addition of a third "mildly emotional" level to traditional neutral and emotional levels enables more fine-tuned detection of anxiety. Controlled experiments have been conducted and found that the three conditions of the emotional Virtual Reality Stroop Task correspond to three distinguishable levels of anxiety. Distinguishing levels of anxiety is a long-standing problem in the field. Additionally, the absence of time pressure and contradictory information in the task helps separate mental load from emotional anxiety.

The second main contribution to the literature is a framework that allows any phobia or anxiety disorder exposure application to be integrated with the seamless feedback system, allowing for transferability and generalisation. The framework relies on implementing certain features, such as changing the exposure levels and standardising exposure levels to apply them to the environment. In this regard, the presented solution differs from many before the current thesis that does not generalise their anxiety feedback solutions to work with various exposure levels. Most previous research focuses on a

single-exposure solution.

These contributions are vital in the progression of the field and add to the efforts of reaching a golden standard for anxiety ground truth determination as well as the generalisation and popularisation of such platforms.

1.3 Publications

Publications in leading Human-Computer Interaction and Computer Science conferences from IEEE and ACM and journals such as PLoS One (Scimago Q1) and MDPI Biosensors (Scimago Q1) have emerged from this PhD. The list of contributions is as follows:

- Deniz Mevlevioğlu, Sabin Tabirca, and David Murphy. Real-time Classification of Anxiety in Virtual Reality Therapy Using Biosensors and a Convolutional Neural Network. 2024. MDPI Biosensors. [MTM24]
- Deniz Mevlevioğlu, Sabin Tabirca, and David Murphy. Anxiety Classification in Virtual Reality Using Biosensors: A Mini Scoping Review. 2023. PLoS One 18 (7). [MTM23a]
- Deniz Mevlevioğlu, Sabin Tabirca, and David Murphy. 2023. Emotional Virtual Reality Stroop Task. IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events. [MTM23b]
- Deniz Mevlevioğlu, Sabin Tabirca, and David Murphy. Emotional Virtual Reality Stroop Task: an Immersive Cognitive Test. 2022. In ACM International Conference on Interactive Media Experiences (IMX' 22). [MTM22a]
- Deniz Mevlevioğlu, Sabin Tabirca, and David Murphy. Real-time Anxiety Prediction in Virtual Reality Therapy. 2022. In Proceedings of the 12th ACM Multimedia Systems Conference (MMSys' 22). [MTM22b]
- Deniz Mevlevioğlu, Sabin Tabirca, and David Murphy. Emotional Virtual Reality Stroop Task: Pilot Design. 2021. In Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology (VRST' 21). [MTM21a]
- Deniz Mevlevioğlu, David Murphy, and Sabin Tabirca. Visual Respiratory Feedback in Virtual Reality Exposure Therapy: A Pilot Study. 2021. In

ACM International Conference on Interactive Media Experiences (IMX '21). [MTM21b]

- Deniz Mevlevioğlu, Sabin Tabirca, and David Murphy. Real-time Anxiety Prediction in Virtual Reality Exposure Therapy. 2021. Adjunct Proceedings of the ACM IMX '21.

[MMT21]

The following open-source dataset has also been a contribution from this PhD:

- Anxiety Classification in Virtual Reality Dataset - OSF.

The contributions emanating from this research, however, were not limited to conference and journal proceedings. This PhD has also resulted in engagement in numerous academic, professional and public outreach activities. These activities can be summarised as follows:

- Poster Contribution at SFI Advance CRT Research Colloquium 2023. A poster entitled "Real-time Anxiety Classification in Virtual Reality Exposure Therapy" was presented at the SFI ADVANCE research colloquium held in Limerick.
- Student Volunteer Chair at ACM International Conference on Interactive Media Experiences (IMX). 2022. A team of student volunteers was directed and monitored to aid hundreds of participants for ACM IMX 2022, the leading international conference for interactive media experiences held in Aveiro, Portugal and virtually.
- Academic Speaker at NBT Berlin: International Forum on Neural Engineering & Brain Technologies. 2022. A presentation entitled "Real-time Anxiety Prediction in Virtual Reality" was delivered to over a hundred people in Berlin.
- Panellist at SFI Advance CRT Research Colloquium 2022. The student contributions section was organised and presented at the SFI ADVANCE research colloquium held in Dublin.
- Poster Contribution at SFI Advance CRT Research Colloquium 2022. A poster entitled "Real-time Anxiety Classification in Virtual Reality Exposure Therapy" was presented at the SFI ADVANCE research colloquium held in Dublin.

1.4 Thesis Structure

The chapters in this thesis are as follows:

- **Chapter 1** outlines and explains the motivation behind the thesis and the scope. It lists the novel contributions to the state of the art and publications emanating from the research and summarises the structure of the thesis.
- **Chapter 2** provides a comprehensive overview of the state of the art informing the study. It details previous research on anxiety disorders, Virtual Reality exposure therapy, wearable sensors to detect stress, and biosensor use within Virtual Reality. It also includes the scoping review entitled "Anxiety Classification in Virtual Reality using Biosensors".
- **Chapter 3** presents the development of a novel Virtual Reality application that serves as a starting point for the Virtual Reality Exposure Therapy framework and provides the ground truth detection and data collection application, the emotional Virtual Reality Stroop Task. This chapter outlines in detail all the steps of developing this environment.
- **Chapter 4** details the experiment design and methods for validating the eVRST environment that has been created. It also shows the preliminary formal data analysis to validate the distinction of each condition in the environment and the appropriateness of further machine learning analysis for deeper analysis.
- **Chapter 5** explains the process of using machine learning to differentiate separate anxiety levels using the eVRST data. The steps to develop the machine learning algorithm include data preparation, feature extraction, feature selection, window selection, model comparison and model validation.
- **Chapter 6** details the development of the VRET framework, including the transferable seamless user anxiety feedback system, the therapist platform to control the environment, and the environments included in the prototype.
- **Chapter 7** summarises the thesis and discusses it in the light of current contributions and implications for future research.

Chapter 2

Scoping Literature Review

This chapter discusses the anxiety classification in VR by providing a background of the problem, summarising years of research leading to the current state of the art, detailing similar work, and identifying gaps and shortcomings in the area. This chapter extends the publication entitled "Anxiety Classification in Virtual Reality Using Biosensors: A Mini Scoping Review", published in PLoS ONE journal in 2023 [MTM23a].

2.1 Background

2.1.1 Definition of Anxiety

APA defines anxiety as an emotional response characterised by negative feelings of worrying and physical symptoms such as higher blood pressure, sweating, dizziness and elevated heart rate [Ame22]. Stress is characterised by a similar emotional and physiological response, however, it is usually caused by an external factor as opposed to a prolonged and internal feeling [Ame19]. Fear is sometimes used interchangeably with anxiety, however, it differs from anxiety for being short-lived and responding to an identifiable threat [Ame19].

This study acknowledges the differences between stress, anxiety and fear. However, due to the similar physiological response of these emotions, they are all included in this study for the sake of identifying physiological systems that can classify and cluster these responses into their respective emotion. These studies generally have a common methodology; thus, it is beneficial to investigate all of them.

2.1.2 Anxiety Disorders

Anxiety disorder is an umbrella term for mental health conditions involving excessive and disproportionate levels of fear or anxiety. Anxiety, although sometimes helpful in responding to dangerous situations, can be detrimental when excessive. There are many different types of anxiety disorders, ranging from generalised anxiety disorder to social anxiety disorder (SAD), from specific phobias to selective mutism [Ame22]. The treatment for anxiety disorders is in response to the particular type of anxiety disorder. The most common treatments would include psychotherapy methods such as cognitive-behavioural therapy (CBT) and medication such as selective serotonin reuptake inhibitors (SSRIs) or serotonin-norepinephrine reuptake inhibitors (SNRIs) [SS22].

Physical symptoms of anxiety disorders can include heart palpitations, shortness of breath and chest pain. Other symptoms can include restlessness, irritability and inability to sleep or concentrate. They result in significant impairments to the patient's life, making it difficult to carry out daily tasks or work. The aetiology of anxiety disorders constitutes a complex interaction including psychosocial factors such as childhood trauma, stress and genetic predisposition [BBW17]. Anxiety disorders are some of the most prevalent mental health conditions globally [ALEL$^+$18, BBW17]. Therefore, screening, diagnosis and treatment of anxiety disorders are crucial to the wellbeing of society.

2.1.2.1 Specific Phobias

Specific phobias refer to the persistent fear and avoidance of a particular stimulus, be it an object or a situation [Ame23b, EBM18]. The lifetime prevalence of specific phobias fluctuates between 1.5% [FAB$^+$04] and 14.4% [KTC01] based on geography and measurement methods, with a median of 7.2% [EBM18]. This variance highlights the importance of contextual factors in understanding and diagnosing specific phobias.

The avoidance of feared stimuli can dampen the effects of the disorder, allowing individuals to function relatively normally. Yet, it remains one of the more persistent and prevalent anxiety disorders [EBM18]. The most common phobias include fear of various animals, such as spiders and snakes, heights, and flying [EBM18]. Each type of phobia can significantly impact an individual's daily life and limit their activities. Despite their prevalence, only 1 in 10 patients receive treatment [EBM18]. This low treatment rate may be due to the stigmatisation of mental health issues or a lack of accessible mental health services.

Due to their persistence and high prevalence, it is vital to study the effects of different treatments and find incentives to make treatment more attractive. Treatments such as CBT can significantly improve the quality of life for those affected by specific phobias [BAM+20].

2.1.2.2 Social Anxiety Disorder

Social anxiety disorder (SAD) is marked by the persistent fear of social situations and performance anxiety [Ame18b, MH13]. Patients generally try to avoid stressful situations, such as speeches or social events [Ame18b]. This avoidance can significantly impair daily functioning and overall quality of life. SAD has a high lifetime prevalence of 12.1% [KBD+05], making it a high-priority challenge in healthcare. The disorder can severely limit an individual's ability to form and maintain social relationships, pursue career opportunities, and engage in everyday activities.

Risk factors include genetic factors, such as behavioural inhibition and withdrawal, and non-shared environmental factors, such as parental influences, peer influences, culture, social outcomes, and adverse life events [SR16]. Genetic factors seem to affect prevalence more commonly in the younger population compared to adults [SR16]. This discrepancy suggests that early intervention could be crucial in mitigating the long-term impact of SAD.

Treatments for SAD typically involve a combination of CBT and, in some cases, medication. Understanding the complex interaction between genetic and environmental factors can help in developing more targeted and effective interventions for persons suffering from social anxiety disorder [MH13].

2.1.2.3 Generalised Anxiety Disorder

Generalised Anxiety Disorder (GAD) is characterised by persistent and excessive worry about a variety of topics. The worrying is hard to control and affects the patient's daily life [Ame18a, TB06]. The lifetime prevalence of GAD is 5.7% [KBD+05]. Chronic anxiety can interfere with daily activities, relationships, and overall wellbeing. The disorder is more prevalent among women, people with lower socio-economic status, and those who experienced childhood adversities [SS15]. These demographic factors suggest that social and economic conditions play a significant role in the development and maintenance of GAD. The genetic factors are moderate, ranging up to 20% among twins [HPM+05]. The rate of cooccurrence in twins indicates that while genetics contribute to the risk of

developing GAD, environmental factors also play a crucial role.

Effective management of GAD often includes a combination of psychotherapy and medication. Ongoing research is focused on identifying the most effective treatment strategies and understanding how various factors contribute to the persistence of GAD [TB06].

2.1.2.4 Panic Disorder

Panic disorder is characterised by regular panic attacks that are accompanied by a fear of another panic attack, the effects of the panic attack, or significant behaviour changes relating to panic attacks [Ame23a, RBCS06]. These attacks can be extremely debilitating, causing individuals to avoid situations where they fear an attack might occur. Panic disorder is often linked to agoraphobia, which is a disorder that marks the avoidance of situations where a panic attack might not be addressed quickly and effectively [Ame23a]. For example, a person with agoraphobia might avoid using public transportation for fear of having a panic attack in an environment where help is not readily available.

The lifetime prevalence of panic disorder is 4.7% [KBD+05]. Risk factors include female gender, genetics, behavioural inhibition, and adverse childhood events [BST+02]. Women are more likely to develop panic disorder, possibly due to hormonal factors or socialisation patterns. Genetic predispositions also play a significant role, suggesting that family history should be considered in assessments and treatment plans. Behavioural inhibition and adverse childhood events can contribute to the development and persistence of panic disorder, highlighting the need for early intervention and support. Treatment typically involves a combination of CBT and medication. Research continues to explore the most effective ways to help individuals manage panic disorder and improve their quality of life [RBCS06].

2.1.3 Anxiety Response

Stress is a useful feeling that can trigger a fight-or-flight response when a person faces danger [McE07]. However, if this response is regularly activated without the presence of a real threat, it might become detrimental to the person's health [AOW+13].

The autonomic nervous system (ANS) controls the bodily responses to fear and anxiety, such as changes in heart rate and respiration [McC07]. The ANS has

two branches: the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). The PNS controls the rest and digest response and how the body acts at rest. The SNS is responsible for the "fight-or-flight" response, which can cause physiological effects of increased heart and breathing rates. It also regulates varying bodily functions such as pupil diameter and gut movement. Anxiety generally involves the activation of SNS; however, it is also common for them to be activated or deactivated concurrently [HRW19].

Physiological outcomes of anxiety depend on several factors, such as the person's age and genetics; as such, these vary from person to person. The short-term response to stress can lead to increased sweating, pupil dilation, heart rate and respiratory rate [Sme10]. The long-term response to stress can lead to increased synthesis of cortisol, which is also called the stress hormone [GRDP+18].

In the existence of a stressor, two systems respond to ensure that the body deals with the stressor efficiently. The first response is from the sympathetic adrenomedullary system (SAM), and it is quick and short-lasting [GRDP+18]. SAM axis works by neurally instigating responses in specific organs with the help of adrenaline produced by stimulation of the hypothalamus [BSSM14]. This results in increased sweating, pupil dilation, and heart rate and breathing rate [Sme10].

The second phase is the response from the Hypothalamic Pituitary adrenal (HPA) axis, which is slower and long-lasting [GRDP+18]. The HPA axis is instigated by the hypothalamus and responds through hormones [BSSM14] and is signified by the synthesis of cortisol.

2.1.3.1 Psychophysiological Markers for Anxiety

Galvanic Skin Response (GSR) measures the change of sweat production in the skin, which is controlled by the SNS. This signal is also referred to as electrodermal activity (EDA). The change in sweat production also affects the electrical conductivity of the skin, which can be measured as a valuable marker for arousal [HRW19]. GSR is a helpful measure of arousal because it is mainly an involuntary process that increases objectiveness. It is a standard measure in studies based on arousal [GP21].

Heart rate levels change based on the activity of both PNS and SNS, with the effect of norepinephrine and acetylcholine [GGG+19]. Increased heart rate points to an overactive PNS, which makes it a straightforward method of

increased anxiety. It can be measured by dividing 60 by the time between R-R peaks in seconds. HR is commonly used in studies with reported positive correlation with stress [GP21]. Heart rate variability (HRV) refers to the time variation between R-R peaks and usually indicates increased SNS activity.

Salivary cortisol levels can indicate the level of activity by the adrenal glands, and a rapid increase in its secretion can indicate that the person is in distress [HRW19].

Body temperature and local skin temperature (SKT), such as fingertip temperature, are also known to change with anxiety. Overall, body temperature is observed to increase under high anxiety. Skin temperature reacts differently based on the location of the measurement. Some studies claim that fingertip temperature decreases under anxiety, and others suggest that facial skin temperature increases.

Brain regions such as the medial prefrontal cortex (MPFC) and hippocampus show a relation between the increased stress stimulus and the inhibition of the HPA axis [Bre04]. As opposed to the MPFC and hippocampus, the amygdala plays a primary role in increasing the HPA axis. Therefore, observing the activity in these regions can provide insights into anxiety levels. There is heterogeneity in results relating to anxiety and brain activity based on the type of anxiety and stressor that has been used in the experiment. However, some brain changes are consistently linked to increased anxiety. These include reduced cerebral blood flow in the orbitofrontal cortex, cerebral blood flow change in the anterior cingulate cortex and deactivation of the hippocampus [DDP09].

2.1.3.2 Behavioural Markers for Anxiety

In addition to physiological changes, increased anxiety can also come with some behavioural effects.

Respiration is one of the most important behavioural markers for anxiety. Breathing rate and depth are known to be related to anxiety, with increased breathing being strongly correlated with increased anxiety [GGG+19]. Fear-potentiated startle (FPS) refers to the startle reflex in response to a fearful stimulus governed by the amygdala.

Voice and speech patterns can show variations under the effect of stress and anxiety [GP21]. One of the most studied components of speech in relation to

anxiety markers would be the pitch of the voice. Typically, a person's speech increases in pitch when they are anxious. Anxiety can also impact patterns in speech, with people emphasising different frequency bands, making these qualities valuable for anxiety detection. Other useful markers for anxiety analysis include excitation, intensity, spectral tilt, gain, jitter, and shimmer [GGG+19].

Movement of the body also changes under high anxiety. An example of this is less movement in the upper body. Frequent and fast head movement is also associated with high anxiety conditions [GGG+19]. Pupil movement is usually associated with changes in light but can also be linked to arousal levels. Increased pupil dilation can be used as a marker for stress [GGG+19]. It is also possible that intense fear can momentarily disrupt breathing. This can lead to irregular patterns of breathing. It is evident that variety in breathing patterns is linked to higher anxiety levels.

2.1.4 Sensor-based Anxiety Measurement

There has been interest in health and mental wellbeing-related areas for methods of objective prediction of anxiety. This can be useful for measuring anxiety in situations where self-reporting is not possible to obtain or is deemed untrustworthy. The use of sensors to measure various markers is common. The selection of sensors depends on a number of factors based on the conditions of the study. In many medical studies where accuracy is vital, it is common to use very precise methods of measurement. Some examples of this would be using fMRI for measurements involving imaging of brain regions. These are considered gold standards, but they can be very limiting for studies that require in-the-wild data or continuous recording of measurements. This study focuses on non-invasive and mobile methods of measurement. A summary of useful non-invasive methods of measuring anxiety can be viewed in Figure 2.1.

A common measurement for getting information on heart activity, including HR and HRV, is through using an electrocardiogram (ECG). ECG records electrical signals from the heart contractions using electrodes placed on the chest. It is very reliable in determining heart-related activity. Due to the direct relation of heart activity with the PNS, this is a commonly used method for anxiety measurements. Another way of measuring heart activity is photoplethysmogram (PPG). PPG measures the pulse rate and pulse rate variability using optical properties of the skin. It is not as reliable as ECG but more cost-efficient and

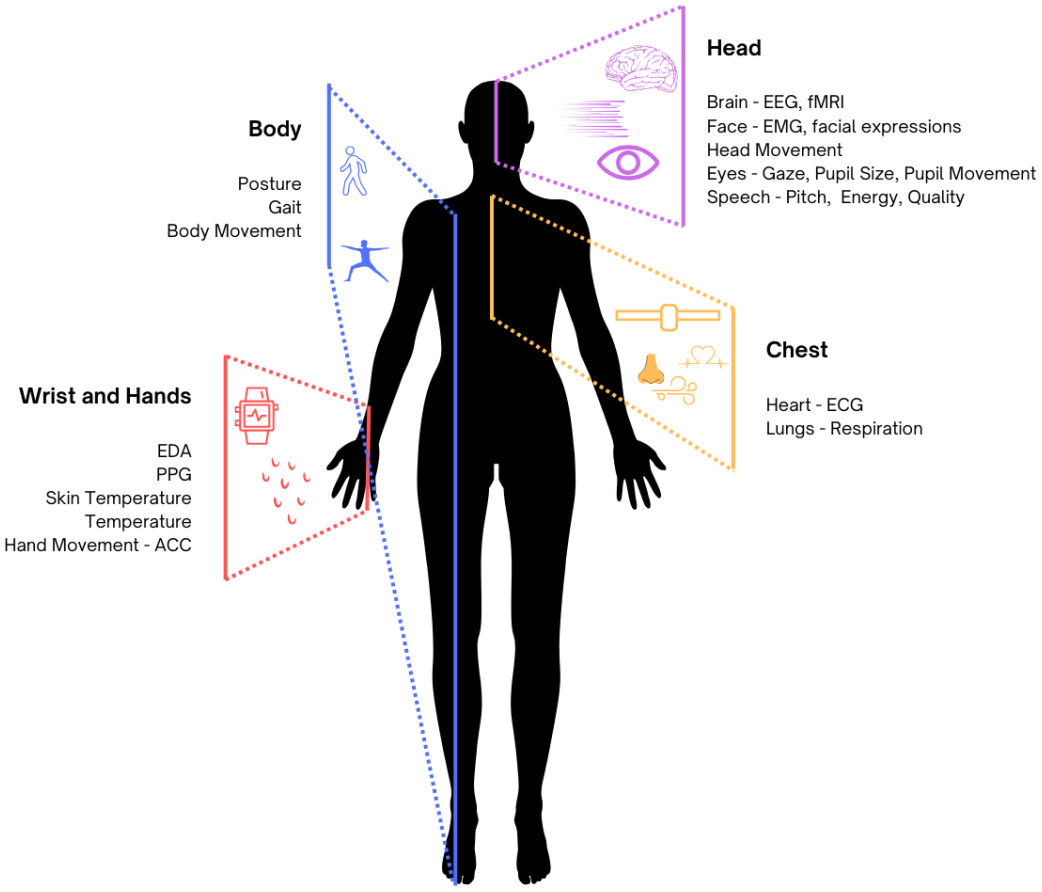


Figure 2.1: Biosensors useful for measuring anxiety

user-friendly, making it a popular research choice. PPG can be obtained from many locations on the body, with popular methods being a wristband or optical pulses attached to earlobes.

EDA measures electrical activity within the skin by recording the electrical resistance between two electrodes, typically placed on two fingers. It can also be obtained from feet. Skin conductivity has two components: phasic and tonic. The tonic component, also referred to as skin conductivity level (SCL), measures the long-term electrical properties of the skin, changing slowly. The phasic component, also called skin conductivity response (SCR), measures peaks in conductivity in response to a stimulus. Two common methods for separating the underlying signals are Continuous Decomposition Analysis (CDA) and Discrete Decomposition Analysis (DDA). The continuous approach deconvolves the GSR signal into tonic and phasic activity by applying inter-impulse fitting of driver data [BK10a]. DDA uses nonnegative deconvolution to decompose the GSR data to SCL and SCR components [BK10b].

An electroencephalogram (EEG) measures the electrical activity in different regions of the brain by using electrodes attached to the scalp. This can be used to infer increases in activity in different regions of the brain, which are associated with the anxiety response. Asymmetry of electrical activity in the left and right lobes of the brain is used to reveal increased emotional arousal [BNK+11]. This can be measured by comparing alpha band activity on the brain's left side to the right. Alpha, beta and theta bands are the most useful measures for detecting anxiety [GP21].

Movement of the body and extremities, such as limbs and the head, can be measured using image-based techniques such as video footage or accelerometers (ACC). Additionally, muscle movement potentials can be measured using electromyography (EMG). An EMG device measures the electrical activity in the muscles by using electrodes attached to the skin surface of a muscle. FPS can be measured through the eyelid movements using EMG [HRW19].

Respiration can be measured through the expansion of the chest cavity or airflow through the nose. To measure chest expansion, a chest band can be used to measure the stretch effect of the chest.

2.1.5 Anxiety Classification using Machine Learning

Studies in anxiety classification use a variety of methods and techniques. Studies in the past generally relied on rule-based systems and formal tests. For example, an increase in skin conductivity would be interpreted as an increase in arousal level. Arousal levels would then be examined with different methods to see if the increase was a positive or negative one. However, with time, a lot of the studies started moving towards using machine learning. Machine learning is a branch of Artificial Intelligence (AI) that uses algorithms to learn information automatically. Regression can be used to predict numerical outcomes on a continuous scale. Classification is used to predict what category a certain data point belongs to. Although there are some numerical systems to grade anxiety, it is difficult to place anxiety on a scale. Therefore, many studies which use machine learning algorithms to predict anxiety use classification. This method of machine learning involves measuring markers for anxiety, such as heart rate or head movement, and using this information to place the user in a predetermined condition. If predicting anxiety on only two levels - either the user is anxious or not- then the classification model would try to place the user in one of two conditions based on the information provided. If the model were

to pick randomly, it would have 50% accuracy. As the number of outcomes increases, the challenge of the classification increases.

Common methods of classification across the literature include Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Decision Trees (DT); including Random Decision Forests (RDF) and Gradient-boosted Trees (GBT), Artificial Neural Networks (ANN), Deep Neural Networks (DNN), and Convolutional Neural Networks (CNN).

2.1.6 Non-VR Systems for Measuring Anxiety

Objective anxiety feedback is useful in many areas, including clinical treatment, patient monitoring, and accessibility assessment. Therefore, there are many biosensor and machine learning based systems outside of a VR environment for measuring anxiety levels. The most popular methods for in-the-wild measurement of anxiety are non-invasive ones that can be comfortably worn throughout the day. Some of the research in the last 20 years can be observed in Table 2.1.

In 2011, de Santos Sierra et al. developed a system that can distinguish between hyperventilation, task preparation, and rest states by using fuzzy logic algorithm [dSSÁCDP11]. Among 80 participants, they were able to distinguish each stage with 99.5% accuracy. They used GSR and HR values to make the distinction.

Castaldo et al. [CXM+16] investigated the effect of exams on the HRV of 42 university students. They were able to differentiate between a rest day after holidays and during a verbal exam in the university with 80% accuracy using an AdaBoosted Decision Tree.

A study by Zhang et al. in 2016 [ZWLH16] examined the GSR levels of participants during a public speaking acticity and while anticipating the public speaking activity. They recruited 22 female participants, and they were able to differentiate between rest and real public speaking with 86.7% accuracy and between anticipatory and real public speaking with 78.8% accuracy.

Vanitha et al. [VK17] produced 89.1% accuracy in differentiating between three different stress levels induced by solving varying difficulty of maths problems. They employed an SVM as a classifying method. They were able to show that EEG is reliable in classifying anxiety. However, they were limited by a small sample size of only 6 participants, which might be difficult to generalise to a wider population.

Table 2.1: Previous non-VR specific research on real-time anxiety prediction

Ref	Sample size	Sensors	Stressor	Model	Output Level	Accuracy
Can et al., 2019	21	ACC, HR, GSR	Contest, Lecture	MLP	3	92.1%
Akmandor et al., 2017	32	GSR, ECG, RESP, BO, BP	Memory Task, Fly Sound, IAPS, Ice Test	SVM	2	95.8%
Vanitha et al., 2017	6	EEG	Maths	SVM	3	89.1%
Chen et al., 2017	14	GSR, ECG, RESP	Driving	ELM	3	99%
Castaldo et al., 2016	42	HRV	Exam	DT	2	80%
Zhang et al., 2016	22	GSR	Public Speaking	BPNN	2	86.7%
de Santos Sierra et al., 2011	80	GSR & HR	HV, TP	Fuzzy Logic	2	99.5 %

Note: ACC: Accelerometer, BO: Blood Oximeter, BP: Blood Pressure, BPNN: Backpropagation Neural Network, DT: Decision Tree, ECG: Electrocardiogram, EDA: Electrodermal Activity, EEG: Electroencephalogram, ELM: Extreme Learning Machine, HR: Heart Rate, HV: Hyperventilation, IAPS: the International Affective Picture System, LDA: Linear Discriminant Analysis, MLP: Multilayer Perceptron, RESP: respiration, SVM: Support Vector Machine, TP: Task Preparation

SoDA, an automatic stress detection and alleviation system, was developed by Akmandor and Jha in 2017 [AJ17]. They used GSR and ECG, Blood Oximeter (BO) and Blood Pressure (BP) to differentiate between stress and no-stress conditions. They employed methods such as a memory test, listening to a buzzing fly, the international affective picture system (IAPS) , and putting one of the participant's hands in a bucket of ice to create stress conditions. They used an SVM to create generalised and individualised models and achieved accuracies of 95.8% and 89.3%, respectively.

Chen et al. used GSR, ECG and respiration to differentiate between rest, highway driving and city driving conditions [CZY+17]. Using an Extreme-Learning Machine (ELM), they were able to achieve accuracies over 99% at the per-drive level and 89% at the cross-drive level.

A study by Can et al. in 2019 [CCEE19] differs from the previously mentioned

studies by virtue of detecting anxiety in real-life settings as opposed to laboratory testing. They recorded the GSR, ECG and ACC values of 21 algorithmic programming contestants for nine days. Using a Multi-layer Perceptron (MLP), they were able to differentiate between lecture days, contest days and rest days with an accuracy of 92.1%.

2.1.7 Exposure Therapy

The roots of exposure therapy stem from a behavioural model of human emotions and anxiety disorders. Traditionally, it was considered that disorders such as phobias stem from negative associations with the feared stimulus through conditioning. In order to minimise this association, certain behavioural strategies involving increased exposure to the stimulus were used.

Systematic desensitisation is one of the oldest forms of exposure therapy. SD involves a hierarchical model of providing increased contact with the stimulus and a positive emotional state [Dav68]. In order to do so, therapists generally used techniques to provide relaxation, such as employing breathing exercises. If the patient becomes anxious during the therapy, the stimulus is eliminated temporarily. The therapy succeeds when the patient is completely relaxed when presented with the stimulus [Rac64].

Visualisation of the feared stimulus was commonly used as low-level exposure before the patient was prepared to face the actual stimulus. It was discovered that using the thought or the imagination of the stimulus contributes to the extinction of fear and negative association [RS02].

Flooding is a precursor to modern exposure therapy methods, and it involves high exposure to a feared stimulus [RMRF00]. There is no graduation of the stimulus, and it moves very quickly. In this method, the avoidance response is prevented. An example of this would be to place a patient with a fear of snakes in a room full of snakes and lock the door. This therapy method ensures that the patient has to confront their fear and is not able to avoid it until they become calm.

Cognitive-behavioural therapy (CBT) focuses on enhancing behavioural exposure therapy methods with cognitive components, including cognitive restructuring and controlled breathing.

Exposure Therapy methods are the most researched type of intervention for anxiety disorders, with a vast number of randomised controlled trials (RCT)

assessing its efficacy. A large body of evidence shows their effectiveness in eliminating symptoms of anxiety disorders in the long term. They are particularly effective in treating GAD, SAD and specific phobias.

2.1.8 Virtual Reality Exposure Therapy

There is a growing body of evidence that points to the efficacy of Virtual Reality Exposure Therapy for the treatment of anxiety disorders, particularly SAD and specific phobias. Although the treatment methods are generally comparable to in vivo treatments, there is no evidence pointing to improved outcomes with the use of VR. They are nonetheless beneficial due to their cost-efficiency and increased reach.

Since the first successful case study conducted by Rothbaum in 1995 [RHK+95], Virtual Reality Exposure Therapy (VRET) is becoming more and more popular by the day [Riv05, WW98]. Using VR as part of exposure therapy has several benefits. First, it is an additional step in systematic desensitisation, which helps adjust the patient and make for a smoother transition between levels [Riv05]. Another significant benefit would be that it does not rely on the patient's imagination to be as vivid as possible to be effective [CWHW09]. In traditional exposure therapy, the user relies on images, photographs and their imagination to simulate the stimulus without the risk of harm caused by elevated responses due to fear [Dav68, Mye06]. Additionally, in a virtual environment, it is straightforward to control the variables and the current exposure and adjust it based on the patient's needs. This can be used to tailor the experience to increase therapy effectiveness.

There have been many studies assessing the effectiveness of VRET in treating different types of anxiety disorders [WL13, OPGP+11]. While the effectiveness of each system varies, the meta-analysis conducted by Opris et al. in 2012 [OPGP+11] over 23 studies with a total of 608 subjects found that VRET is significantly more effective compared to waitlist control and it produces similar results to in vivo exposure therapy. The study also found that VRET has a powerful impact on patients' daily lives, with persistent results over time. This evidence strongly points towards VRET being a viable treatment for anxiety disorders.

It should, however, be noted that immersion plays a critical role in the effectiveness of VRET [HRK+94]. In a controlled study conducted by Pallavicini et al. in 2013, [PCR+13], it was revealed that without a sense of presence, VR

does not perform as well as stressors in real life to produce a negative emotional response. Without the feeling of presence, VR becomes comparable to media such as videos or images, which are much cheaper to administer. So, presence is an essential aspect of any VRET. Although different studies point to different results, in the study by Krijn et al. [KEB+04], it was revealed that there were no differences in the effectiveness of VRET done using head-mounted displays (HMD) and computer automatic virtual environments (CAVE). CAVE leads to a much higher presence. This might show that the sense of presence using an HMD is already sufficient. In Hodges' study [HRK+94], it was shown that the sense of presence can be achieved without being a perfect replication of real life. Also, the users being exposed to the actual scene during the experience modelled after did not affect the users' sense of presence.

In addition to being successful in producing the effects that it is trying to replicate for people who have an anxiety disorder, in a study by Schuemie et al. [SBD+00], it was discovered that regardless of the user's general fear of heights, being higher in the virtual environment always corresponded with elevated levels of anxiety. This is presumably due to heights innately being assessed as dangerous by humans; however, acrophobia would be defined as an irrational fear of heights beyond the average person's fear [CW10].

One recent control study by Anderson et al. [APE+13] investigating the effectiveness of VRET in treating social anxiety disorder compared to in vivo treatments found no significant differences in any outcome measure and no difference in partial or full remission rates. This has been maintained over a one-year follow-up.

Similarly, a study by Bouchard et al. [BDR+17] found improvements to all outcome measures compared to waitlist controls in both in vivo exposure therapy and VRET. Furthermore, there were improvements above in vivo in the post-treatment primary outcome and one secondary outcome. This improvement persisted over a six-month period. Psychiatrists reported that using VRET was more practical than completing it in vivo [BDR+17].

Meta-analyses show no significant improvement in therapy outcomes for VRET over in vivo [CSLC+19a, WKM19]. However, they do show an improvement over waitlist and placebo controls [CSLC+19a].

Although the treatment methods are generally comparable to in vivo treatments, no evidence points to improved outcomes using Virtual Reality.

They are nonetheless beneficial due to their cost-efficiency [DBH+15] and practicality [BDR+17]. An experiment by Garcia-Galacios et al. [GPHKS+01] shows that 81% of students preferred VRET to in vivo treatment. However, based on the meta-analysis by Benbow and Anderson [BA19], the dropout rates did not show any significant difference.

2.1.8.1 Limitations of Virtual Reality

Although VR provides cost-efficiency and accessibility, there are some limitations to using VR for mental health treatments. One of the main limitations of VR for these applications is its inaccessibility to people with vision problems. While near-sighted people can easily use the platform, far-sighted people, people with poorer vision in one eye, and people with limited vision might not be able to fully benefit from the VR environment. Some HMDs allow for the use of glasses; however, it is often not comfortable. There are lenses that can be attached to the HMD for personal devices, but this limits the treatment for people who can afford such systems.

A major concern is motion sickness derived from VR, also called cyber sickness. The leading theory of motion sickness suggests that the main influence is postural instability, meaning the lack of autonomy over the maintenance of one's posture [RS91]. According to recent studies using hardware, from 30% to 80% of users experience motion sickness in VR [RO16]. It should be noted that measuring cybersickness is challenging due to the absence of a gold standard and the bias of self-report measures [RO16]. To be able to reduce cybersickness outcomes, it is important to determine its principal causes. A recent review by Chang et al. [CKY20] breaks down the factors into three categories: hardware, content and human. With the improvement of HMDs over time, motion sickness has been steadily decreasing. However, this decrease has not been enough to eliminate the problem [RO16]. Some hardware factors that affect motion sickness include field-of-view (FOV), latency and flicker [CKY20]. The content-related factors include visual movement, controls, the realism of the graphics, content FOV and experience duration [CKY20, SSB+20]. Some examples of this might include discrepancy between visual movement and physical movement, objects in VR not moving as would be observed in real life and increased fidelity in VR, increasing immersion and possibly the discrepancy of expected movement [CKY20]. Motion sickness can be observed in experiences as short as ten minutes and is positively correlated with the length of the VR experience [SSB+20]. Based on a review by Saredakis et al. in 2020 [SSB+20],

the type of VR content that resulted in the highest level of motion sickness were games with high fidelity and high interaction level, followed by 360 videos. The content that caused a low amount of motion sickness was minimalistic environments with simple shapes and interactions and scenic environments with little to no interaction. Based on limited evidence, the motion sickness levels were lower for those over the age of 35 [SSB+20]. However, more research into the effects of motion sickness in older adults would be needed to draw more concrete conclusions.

Another significant limitation is the accessibility of new technology in older adults. According to a meta-analysis by Hauk et al. in 2018, [HHK18], technology acceptance negatively correlates to age. A study by Huygelier et al. in 2019 [HSVE+19] investigated the attitude change in 76 adults aged 57 to 94. Although they initially had neutral attitudes towards VR, after HMD exposure, this attitude increased to a positive one [HSVE+19]. The participants also reported minimal motion sickness. This demonstrates that the challenge of older adults' attitudes towards this technology can be overcome with exposure and information.

Although limitations for accessibility to VR exist, so do strategies to overcome these limitations. When designing VR systems for therapy use, it is essential to account for the needs of the population, including people with poor vision, old age and motion sickness. With thoughtful design choices, the negative effects of VR can be minimised.

2.1.8.2 Biosensors in Virtual Reality

Accessibility and the helpfulness of VR led many researchers to seek ways to improve VR therapy. One such way is tailoring the therapy environment to the patient's needs. To do this, however, continuous feedback from the patient is required. Self-report feedback can be unreliable and challenging to obtain during a VR session. Interruptions to exposure to fill questionnaires can take a toll on the immersiveness of the experience and increase the likelihood of dropout [BA19]. Therefore, providing objective and non-intrusive feedback during a VR therapy experience is crucial.

Extensive research shows that it is possible to measure and categorise distinct anxiety levels using biosensors [GP21]. However, not all devices used for measuring physiological signals are suitable for use in VR. The highest priority in VR is to allow for user mobility and comfort so users can explore and

experience the environment at their own pace [YK19]. However, a lot of on-body sensors, such as ECG, PPG and EEG, work best when the user is stationary [ASR+03]. Due to the noise introduced by movement and the electrical interference of the HMD, it is important to use strategies such as filtering, multi-modality and normalising to offset the negative effects. While many preprocessing methods on physiological data ensure the highest quality results, not all are viable for real-time use.

Common approaches to measuring anxiety in VR involve using physiological and biological markers for anxiety [GGG+19]. Standard physiological measures include HR, EDA, electrical brain activity and SKT. Common behavioural measures include head movement, muscle movement, eye movement and respiration (RESP) [GP21]. Numerous studies combine machine learning methods and on-body biological sensors for optimal, accurate and objective feedback on anxiety levels [GPT+12, vDML19, SC15].

Due to its physiological response, the anxiety level is strongly and positively correlated with both SCL and HR, resulting in these measures being among the most popular for anxiety classification [GGG+19]. Other physiological measures that are correlated with anxiety include SKT, blood volume pressure (BVP), electrical brain activity and electrical muscle response [GGG+19]. Compared to heart rate and skin conductance, the interaction between these measures and anxiety is more complicated, however, they are evidenced to be useful measures.

Anxiety also results in some behavioural responses, which can be detected through behavioural measures such as head movement, eye movement and respiration. Evidence suggests that this information also has a relation with anxiety, making them useful metrics for classification [AEM22, GP21].

Salivary cortisol levels are positively correlated with anxiety and stress [GGG+19, GRDP+18]. However, the methods for determining levels of cortisol in saliva are generally invasive, which might not be optimal for an immersive VR experience.

Prior studies made use of psychophysiological measures and machine learning models to classify predefined levels of anxiety, such as predicting if the person is currently experiencing relaxation, mild anxiety or severe anxiety [GP21].

Over the last decades, there has been increasing research in classifying anxiety using machine learning and sensors [GGG+19]. Unfortunately, many of the systems used to detect anxiety using sensors limit the mobility of the user

[GP21], which renders them unsuitable for VR treatments as movement plays a very important role in presence within VR [SSMM98, REG+19]. This gap indicates a need for investigating objective measures using physiological and behavioural information that allows for higher mobility so it can be effectively used with VR.

2.2 Scoping Review

VR has become more and more popular by the day, while the use of VR for medicinal and therapeutical reasons is almost commonplace. While VR can make valuable contributions to the health and wellbeing of society, it is important to research and monitor the effects it has on individuals.

One such area that VR has been evidenced to be very effective in is fighting anxiety disorders [Riv03]. As more proof of the efficiency of VR therapy is unearthed, the study of feedback collection within Virtual Environments (VE) becomes more important. One such feedback is how anxious the user feels within such an environment.

Real-time anxiety classification within VR enables a range of possibilities for VR therapy, supporting decision-making in VR and providing feedback on the effectiveness of the intervention. It, unfortunately, comes with a set of challenges, such as the bias introduced by self-assessing anxiety [BBM01], not breaking the immersion in VR and the technical difficulty of systems that measure physiological signals while the user is mobile [GGG+19].

The current review is a mini-scoping review using PRISMA-ScR guidelines with the topic of anxiety classification in VR using biosensors [PMB+21].

2.2.1 Application areas

VR is commonly used for therapy, meditation and medicinal purposes [EM21, Riv03]. Systems that incorporate wearable sensors are not commercially widespread, however, the research in this area is currently extending due to its many areas of application [CXC19]. Using VR for health care can increase cost efficiency and provide support when in vivo options are difficult to reach, and it can also help reduce the load of health care systems [MG16, RHK97]; therefore, VR treatment methods for psychological disorders are becoming increasingly popular while the scientific evidence points to similar effectiveness when

compared to in vivo treatment methods [EM21, KEOB04].

As VR therapy has clear benefits of cost-efficiency and availability, it is important for the wellbeing of society to research it as a medium [MG16, Riv03]. Although it might not be accessible for everyone at the moment due to problems with adjustment to the virtual world [HSVE+19], possible visual difficulties for some users [LIF+09], and motion sickness [CKY20, NP02], it is still an important addition to current therapy methods [LYS+17]. Feedback from the patient is vital for the improvement of VR therapy. While the most direct approach to this is asking the user, this method can either be disruptive or unhelpful. Also, it is challenging to get self-report measures immediately, and it can be tricky to self-assess retrospectively as the user might have already forgotten how they were feeling at the time. This is why it is crucial to have a real-time solution. Physiological signals using wearable sensors provide a way of receiving feedback at the moment, objectively, and without interruptions to the experience.

A promising application of anxiety detection in VR is its use in conjunction with VRET [LKH+21]. VRET is a therapy method to treat certain anxiety disorders by allowing the user to face their fears in the safety of the provided virtual environments [CSLC+19b]. As every person is unique and responds to therapy differently, it is important to predict the anxiety level of the user to make better-informed decisions on the exposure adjustment.

In addition to its medicinal and therapeutical applications, real-time anxiety classification in VR is useful for any type of application that uses feedback. How users are feeling at the moment can be used to adjust the content, make improvements to software, and change settings. One example of its use is providing insight into user state while performing challenging tasks [TGGC15]. It is also interesting to look at it for entertainment purposes, using the classification of anxiety to adjust difficulty/atmosphere in immersive VR games [DC07].

2.2.2 Research Questions

This review sets out to answer the following research questions:

- RQ1: What are the most common measures for anxiety classification, and what are their advantages and disadvantages?
- RQ2: What are the current ground truth establishment methods used to

classify anxiety, and what are their advantages and disadvantages?

- RQ3: What are the classification methods most commonly used, and what are their advantages and disadvantages?
- RQ4: What are the gaps in the literature relating to real-time anxiety prediction in Virtual Reality?

2.3 Materials and Methods

Inclusion and exclusion criteria can be viewed in Table 2.2.

Table 2.2: Inclusion and exclusion criteria

Inclusion Criteria	**Exclusion Criteria**
The study was conducted in a VR environment	The anxiety classification is not done in a VR environment
The experiment examines user anxiety in real-time using classification models	The anxiety classification cannot be applied in real-time
The paper is peer-reviewed short or full paper	The study is not a full or short paper or not peer-reviewed
The study was published in or after 2010	The study was published before 2010

IEEE Xplore, ACM Digital Library, SCOPUS, and Web of Science were searched for eligible studies. The following strategy was used to search each database; title, abstract or topic, including “Virtual Reality” or “VR” and “anxiety” or “stress” and “recogn*” or “detect*” or “predict*” or “classif*” or “discriminat*” or “identif*”, filtered to only include publications published between 2010 and 2022. Supplementary sources were scanned using references from identified papers, past reviews with similar titles and Google Scholar. After records were identified, duplicates were removed. The resulting records’ titles and abstracts were then screened by a single reviewer for relevancy. The relevant papers were fully examined for eligibility.

The review was not registered, and a protocol was not prepared. The completed PRISMA-ScR checklist can be found in Appendix A. A data extraction form was created and pilot tested on five studies. The data were extracted from the studies using the data extraction form. The full data extraction form can be viewed in Appendix B. The extracted data can be viewed in S1 Table. The following are the items in the data extraction form:

- DI1: Reference (First author / Year / Citation)
- DI2: Sample size
- DI3: Type(s) of physiological or behavioural measurements (heart rate, eye movement, etc.) used
- DI4: Ground truth establishment method used
- DI5: Classification model that achieved the highest accuracy
- DI6: Numbers of outputs for the classification
- DI7: Quality score
- DI8: Highest accuracy

The quality assessment scale was modified from the [Cri19] qualitative study checklist and pilot-tested on five studies. It has eight items, with possible answers of yes (1 point), no (0 points) and partially (0.5 points). The possible quality score ranges from 0-8. The complete checklist can be viewed in Appendix C.

2.4 Results

A summary of the search results is presented in Figure 2.2.

Sixteen studies were excluded because they were found to be ineligible upon further investigation [CCN+12, GLC+13, HMPM14a, HMPM14b, HMP+13, IRN+20, JKJ20, LSL+17, KAK21, PSN18, PVVBLGM21, SGT20, ZTXS19]. Exclusion Criteria 2 (classification does not take place in a VR environment) accounted for nine of these exclusions [GLC+13, HMPM14a, HMPM14b, HMP+13, LSL+17, PSN18, PVVBLGM21, RSFAE+12, SGT20], and Exclusion Criteria 3 (no classification that can be applied in real-time) accounted for seven [ADT21, CCN+12, HEK20, IRN+20, JKJ20, KAK21, ZTXS19]. The studies by Handouzi et al. [HMPM14a, HMPM14b, HMP+13] state that their studies take place in a VR environment, but the content is displayed on a 2D screen while the participant is asked not to move their limbs, so these papers were deemed ineligible due to not being applicable to VR. No papers were identified through supplementary sources.

The detailed quality scores of each study can be viewed in Figure 2.3.

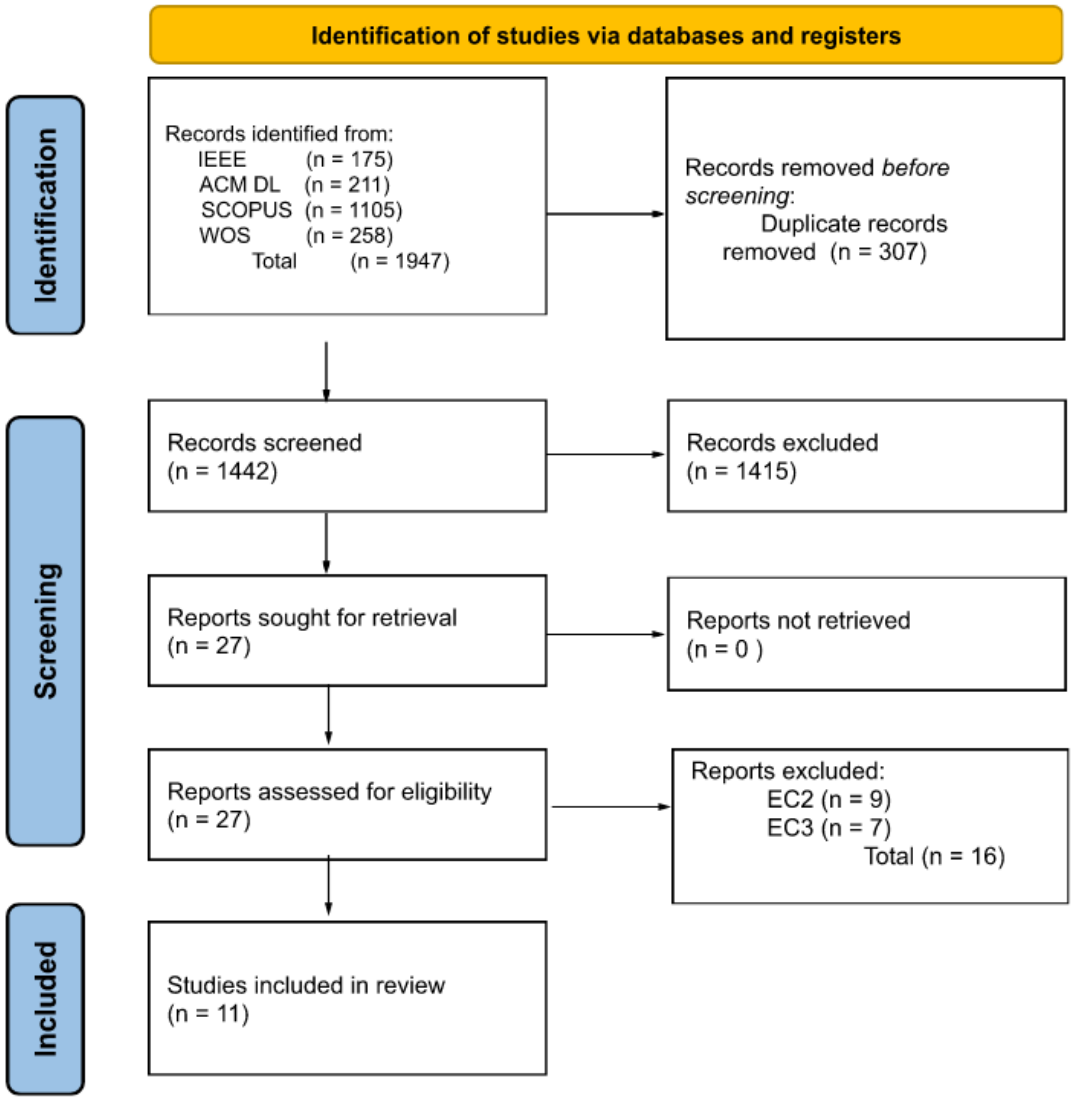


Figure 2.2: The search results. WOS: Web of Science [PMB+21]

A summary of the data extraction from eligible papers is presented in Table 2.3. Most of the studies achieved quality scores over 6, with the only exception being the study by [KHL+20]. When interpreting the table of results, it should be kept in mind that the accuracy of the models cannot be compared directly due to differences between studies in terms of what they classify as anxiety.

2.4.1 Physiological and Behavioural Measures

There was a large variety of measures used between studies. Behavioural measures used were respiration (RESP) [KHL+20, TCT+15, WCL+10], head movement [KHBK20] and eye movement [KSZ21]. Physiological measures used

Table 2.3: Previous research on real-time anxiety prediction in VR

Reference	Sample size	Features	Ground truth	Model	Output Levels	Quality Score	Accuracy
[BFF21]	8	HRV (ECG)	VR Content	LSTM	2, 3, 4	6.5	90.5% (2-Level), 67.5% (3-Level), 58.8% (4-Level), 30s
[BMM+20]	4	PPG, EDA, EEG	SUDS	ANN	11, 4, 2	6.5	78.3% (2-Level), 38.8% (4-Level), 26.5% (11-Level)
[CHO+17]	12	PPG, EDA, SKT	VR Content, STAI-Y1	K-ELM	3	6	96.3%
[HCOL17]	6	ECG, PPG	VR Content	LDA	3	6.5	79%
[KSZ21]	28	EEG, EMDR	SCWT	MLP	2	8	96.42%
[KHL+20]	1 (trained on 40)	ECG, EDA, EMG, EEG, RESP, ACC, TMP	VR Content	SVM	2	4.5	82%
[KHBK20]	73	Head movement	Card Sort, COMP-TRACK, Movies	SVM	2	7	75%
[PPM+20]	7	EDA, PPG	SUDS	Regression with thresholds	2, 3	8	94.3% (2-Level), 92.4% (3-level)
[TCT+15]	20	ECG, RESP	STAI-Y1	Neuro-fuzzy	4	7	83%
[vDML19]	30	BVP, EDA, SKT	VR Content, SUDS	SVM	4	8	86.3%
[WCL+10]	19	EDA, RESP, ECG, EEG	SCWT	SVM	3	7	84%

Note: ACC: Accelerometer, BVP: Blood Volume Pressure, ECG: Electrocardiogram, EDA: Electrodermal Activity, EEG: Electroencephalogram, EMDR: Eye Movement Desensitisation and Reprocessing, EMG: Electromyogram, HRV: Heart rate variability, PPG: Photoplethysmogram, RESP: respiration, SKT: Skin temperature, TMP: Temperature, SCWT: Stroop Colour-Word Task, SUDS: Subjective Units of Distress Scale, STAI: State-Trait Anxiety Inventory, LSTM: Long Short Term Memory, ANN: Artificial Neural Network, K-ELM: Kernel Extreme Learning Machine, LDA: Linear Discriminant Analysis, MLP: Multi-layer Perceptron, SVM: Support Vector Machine, CNN: Convolutional Neural Network

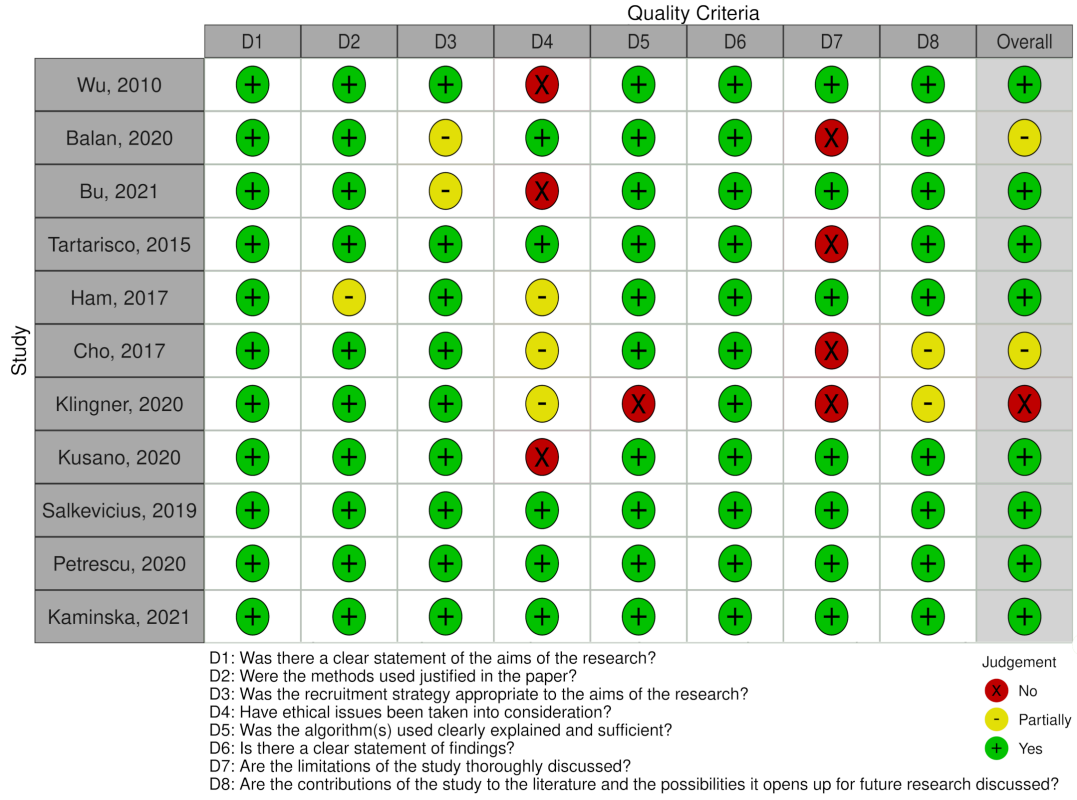


Figure 2.3: Quality scores of each study [MH20]

by high-quality studies were electrodermal activity (EDA) [BMM+20, CHO+17, PPM+20, vDML19, WCL+10], heart rate (HR) and heart rate variability (HRV) measured by electrocardiography (ECG) [BFF21, HCOL17, KHL+20, TCT+15, WCL+10] and pulse rate (PRV) and pulse rate variability (PRV) measured by photoplethysmography (PPG) [CHO+17, HCOL17, PPM+20], brain electrical activity measured by an electroencephalogram (EEG) [BMM+20, KSZ21, KHL+20, WCL+10] and skin temperature (SKT) [CHO+17, vDML19]. [KHL+20] also stated that they used electromyography (EMG), temperature (TMP) and accelerometer (ACC). However, there was no information provided about these sensors, and how and why they were used.

Some of these studies combined behavioural and physiological measurements [KSZ21, KHL+20, TCT+15, WCL+10] while others focused on physiological measures alone. [KHBK20] used only head movement to classify anxiety levels.

As stated before, one of the biggest challenges in deciding the measures is getting objective and high-quality information without causing discomfort in

the VR environment, which can have a toll on immersion. However, it is hard to interpret physiological information from biosensors, due to the difficulty of producing noise-free data and the methods of removing noise in real-time being limited. As a result of this, many studies [BMM+20, CHO+17, HCOL17, KSZ21, KHL+20, PPM+20, TCT+15, vDML19, WCL+10] chose to use multi-modal systems incorporating a range of measures to compensate for the accuracy loss in the case of receiving low-quality signal from any one sensor. It can be argued, however, that increasing the number of biosensors can decrease the sense of presence within VR and that sense of presence is important for VR therapy [KEOB04]. Thus, some studies chose to limit the number of measures to one or two measures [BFF21, HCOL17, KSZ21, KHBK20, PPM+20, TCT+15].

2.4.2 Ground Truth

Ground truth refers to the point of reference used to differentiate between anxiety and no-anxiety states within the experiments. Establishing a ground truth of anxiety is a difficult procedure due to the difficulty of assessing anxiety itself [LZZJ05]. In literature, different approaches have been taken in the establishment of ground truth.

Some studies use self-report measures or questionnaires such as subjective units of distress scale (SUDS) or state-trait anxiety inventory (STAI) [BMM+20, CHO+17, TCT+15, vDML19]. This reduces the ambiguity of what is being detected in the environment, which means there are fewer assumptions. However, if used alone, this method can lead to bias due to personal differences between participants and the difficulty of arbitrarily describing one's own anxiety levels.

Other studies use well-established stress-inducing cognitive tasks such as Stroop Colour-Word Task (SCWT) or card sort task [KSZ21, KHBK20, WCL+10]. These types of cognitive tasks have been widely used in literature and have been validated to cause certain stress reactions. Therefore, some studies use them to objectively label levels of anxiety, reducing the subjectivity of the measure.

Some studies preferred to make their own relaxing and anxious VR environments and validated these to be used as ground truth in their studies[BFF21, HCOL17, KHL+20]. The advantage of this approach is its applicability to its use case. However, studies that only use this approach introduce a risk of bias to their results due to the difficulty of validating their results.

2.4.3 Classification Models

Studies included in this review tried different features and models to come up with the highest accuracy solution (Table 2.3). Not all of the included studies aimed for real-time classification, so for some studies, high performance is not a priority. Given the nature of the features, some studies succeeded in reaching optimal accuracy using relatively simple models such as Linear Discriminant Analysis (LDA) [HCOL17] and Support Vector Machines (SVM) [vDML19, WCL+10]. [PPM+20] found better results using a Convolutional Neural Network (CNN). Some studies found high-accuracy results without compromising performance by using feedforward neural networks such as Kernel Extreme Learning Machine (K-ELM) [CHO+17] and Neuro-fuzzy systems [TCT+15]. However, it should be kept in mind that there are other effects in play, such as biosensors and stressors, and the accuracy of the described models can only be evaluated within their studies.

Features used in the evaluated studies heavily depend on the type of measures they were using. They are commonly grouped into time domain, frequency domain, and non-linear measures. Common statistical features used across most signals were mean, average, minimum, maximum, standard deviation and variance [CHO+17, HCOL17, KSZ21, PPM+20, vDML19, WCL+10]. Common features used relating to ECG and PPG were Inter-beat Intervals (IBI) and Normal-to-Normal (NN) average and standard deviations in the time domain, and low and high-frequency component averages in the frequency domain [CHO+17, HCOL17, PPM+20, WCL+10]. For EDA or GSR, these included event-related responses using peak detection such as number of peaks, peak frequency, average peak amplitude, and maximum peak amplitude [CHO+17, PPM+20, vDML19]. Approximate entropy and sample entropy were non-linear features used to quantify regularity and predictability [CHO+17, HCOL17].

Preparation for physiological and behavioural data varies. For information such as ECG signals and EEG signals, it is very common to use bandpass filters to avoid using artefacts introduced from movement and electrical interference [CHO+17, HCOL17, KSZ21, WCL+10, vDML19, PPM+20]. Normalisation was generally applied in studies that use ANNs and SVMs [vDML19, PPM+20, WCL+10]. Additionally, Petrescu et al. [PPM+20] applied down-sampling because their EDA signal's sampling rate was higher than that imposed by its bandwidth. For EEG, the data was split into separate

wavelengths of delta, theta, alpha, beta, low beta, high beta and gamma [KSZ21, WCL+10]. Wu et al. [WCL+10] focussed specifically on alpha, beta and theta. Independent Component Analysis (ICA) was used by Kaminska et al. [KSZ21] for processing EEG data to remove unwanted noise.

The studies used various feature selection methods to simplify the models after extracting a high number of potential features. Most of the studies in this review used different methods, including Sequential Forward Selection (SFS) [WCL+10], Pearson correlation [PPM+20], ReliefF and Davies-Bouldin cluster evaluation index [TCT+15].

Time windows were relatively small across most studies. Eight of the studies ranged from 3s [KSZ21, WCL+10] to 30s [BMM+20] window durations. The study by Bu et al. [BFF21] featured time windows of five minutes. Only three studies used windows longer than 6s [BMM+20, KHBK20, vDML19]. Two studies did not specify their time window [HCOL17, KHL+20]. As the time windows get bigger, it is easier to make accurate estimations. However, the resulting models might end up being less usable by making the real-time estimations too vague. So, a good balance between accuracy and using smaller time windows is important.

The number of outputs that were classified in these models ranged from two to eleven across all studies, with the most common being two. Many studies used two-output models (no anxiety vs anxiety) because it can be difficult to define and detect increasing levels of anxiety, increasing the challenge of labelling data for classification. However, some studies found success with increased levels of outputs, giving more detailed information regarding the anxiety state of the user [CHO+17, PPM+20, vDML19].

It is hard to determine here what the "acceptable" accuracy is due to the differences and difficulties in defining different anxiety states. Across the studies reviewed, accuracy for two output models ranged between 75% [KHBK20] and 96.4% [KSZ21]; accuracy for three output models ranged between 67.5% [BFF21] and 96.3% [CHO+17]; accuracy for four output models ranged between 38.8% [BMM+20] and 86.3% [vDML19]. In the study by [BMM+20], the accuracy results are compared to the chance of getting the correct results when making a random guess (50% for 2-choice, 25% for 4-choice), however, it is hard to determine how much practical use can come out from such a comparison. Higher accuracy is desirable as low-accuracy models will not be reliable, however, the practical use of the model is important as well. If what is being

predicted does not prove useful in a VR application, the accuracy of the prediction means very little. Furthermore, user comfort must not be compromised when increasing classification accuracy.

2.4.4 Gaps in the Literature

It is difficult to ascertain whether there are enough studies in this specific area yet to conclude the usefulness of anxiety classification systems. The up-to-date studies, despite their problems, show promise for the future. However, there is a big lack of in-the-wild solutions, meaning that most studies were conducted in labs under controlled conditions. Furthermore, in many of the studies, data collection is only done once, and the results are validated in past data and not in real time. Although varying between studies, the accuracy of classification models approaches useful levels. While these results are useful in their contained studies and VR environments, few solutions have been tested across different environments and samples. This makes the generalisability very difficult to assess. Furthermore, there are gaps in the application of these systems to clinical trials, even though many of them are developed to aid therapy.

2.5 Discussion

There are many different measures used in anxiety classification in VR, behavioural or physiological. Many studies used heart rate and electrodermal activity to detect anxiety [BMM+20, CHO+17, KHL+20, PPM+20, WCL+10]. From the behavioural measures, respiration was the most prevalent [KHL+20, TCT+15, WCL+10]. Many high-quality studies were able to achieve promising results by combining several measures to strengthen the reliability of their system [KSZ21, TCT+15, WCL+10]. This might suggest that it is a good idea for future studies to use at least two different measures while classifying anxiety in VR.

Current literature has no agreement regarding important issues that lie in the foundation of the problem of anxiety classification in VR. One of the key issues is the definition of anxiety and relaxation itself. It is understandable that studies measure different levels of anxiety, however, these levels are not linear or easily distinguishable between the studies. What one study describes as the absence of anxiety, another might describe as mild anxiety. As such, it is difficult to answer whether it is possible to accurately classify anxiety in VR.

Certainly, there were many promising results and high-quality research when it came to achieving high internal accuracy [KSZ21, PPM+20, vDML19]. However, the question remains whether it is possible to claim that the anxiety measured here is generalisable and valid. For this to be possible, there needs to be some unification of accepted standards. Thus, working on generalisable standards for anxiety measures must be central for future studies.

2.5.1 Limitations

The studies included in this review had several important limitations. Some studies poorly described their samples and methods [BFF21, KHL+20] which made it difficult to extract meaningful information from these studies. Many of the studies had sample sizes below 10, which makes interpretation of the results difficult [BFF21, BMM+20, HCOL17, PPM+20]. Furthermore, most studies were done in universities and had a young population, with some samples' mean age being under 30 [CHO+17, HCOL17, vDML19]. This is problematic due to two reasons. First, the results can only be interpreted for younger people, which means that the prediction might not be accurate when administered to the general population. Second, it is hard to determine the problems in the system when used by older people. Some studies make the inference that their results show sufficient success in predicting accuracy with little to no evidence [BMM+20, KHL+20].

This study used single screening, which, based on research, might introduce some errors in abstract screening and increase the risk of missing eligible studies [WKS+19, GAT+20]. Also, the study quality criteria creation, rating of study quality and data extraction were conducted by a single reviewer and later verified by a second reviewer. This might have introduced a risk of bias to the study. However, these limitations are estimated to not have majorly impacted the conclusion of the study.

2.5.2 Future Directions

This field is still relatively new, but it enjoys consistent attention and is steadily progressing. Although the direction of research here may be promising, there is a need for a lot more research before it can be used in daily life.

First of all, there need to be a lot more studies on best practices when it comes to establishing ground truth. Studies that test ground truth detection methods

over different applications and settings are vital for the validation of anxiety levels. Without this, it is very difficult to ascertain what is actually being measured.

Second, a lot of the studies in the literature thus far focus on laboratory settings and are contained within one application. There is close to no reproducing of the results, which leads to questions about the validity of the results when it comes to wider applications. The variability in the ground truth, VR content, and measures used make it difficult to compare results directly. Thus, it is important to reproduce previous work using the same methods on different samples.

Another area that needs attention is the application of these systems to clinical trials to validate whether they are useful in their application areas. In-the-wild testing is required to further validate the accuracy of the systems, as well as their usability.

2.6 Conclusion

This review shows that it is possible to consistently measure the anxiety state of the user for the purpose of research. The use of anxiety classification within VR enables developers to further tailor experiences based on user needs, possibly increasing the effectiveness of therapeutical or entertainment VR applications and potentially decreasing dropout rates for therapy. It is vital for future research to study the benefits of anxiety classification in such applications, for instance, the effectiveness of VRET enhanced by real-time anxiety classification as opposed to traditional systems. To study such an effect, longitudinal controlled trials are of priority.

Based on the findings of this review, future studies should prioritise methods and report them very precisely. This is vital for deriving meaningful information from machine learning results. Although there are many studies that work on perfecting the efficiency and performance of algorithms, there are very few studies focussing on establishing standardised techniques for measuring the ground truth of anxiety and anxiety-inducing methods. In many studies, it is difficult to discern cognitive load and mental task load from emotional anxiety or fear.

Transferability is also an uncommon characteristic of many current systems. Presently, many prototypical and pilot studies with anxiety prediction systems

are only suitable for the research and application the researchers use. In many studies, the tools used are not commercially available or widespread. The systems that can be used by a wider audience and various applications would increase the reach of such research.

This chapter has summarised the current advancements in anxiety classification using biosensors and detailed the concepts that will be used throughout the thesis. The next chapter introduces and describes the development of the VR environment that will be utilised in the investigation of anxiety ground truth.

Chapter 3

eVRST: Emotional Virtual Reality Stroop Task Development

This chapter describes the steps taken and the decision-making for creating the emotional Virtual Reality Stroop Task (eVRST) application. eVRST is the Virtual Reality version of the well-established emotional Stroop task. It is an immersive multi-modal application of the real-life Stroop task, simultaneously recording physiological information for later analysis. With this application, the aim is to tackle the long-standing problem of ground truth detection of anxiety within VR. This chapter expands the publication entitled "Emotional Virtual Reality Stroop Task: an Immersive Cognitive Test ", published in ACM International Conference on Interactive Media Experiences in 2022 [MTM22a].

3.1 Background

As detailed in the literature review chapter, recognising a user's emotional state within VR is a long-standing challenge. Accurate and convenient classification of the user state may enable the collection of objective feedback from the user [GGG+19]. This can be useful for various applications where feedback is valued, but distractions are discouraged [RMC+07]. Examples include games, immersive activities, and therapeutical applications [RMC+07]. The so-called "affective computing" is currently significantly impacting VR applications and VR research [DAP+15].

As the primary goal of this thesis is to accurately predict and classify user anxiety to support decision-making in VR therapy, anxiety will be the main

focus in this chapter. In this chapter, anxiety will be predicted using high arousal and negative liking. Anxiety can be defined as the feeling of discomfort that a person feels regardless of the presence of a stressing stimulus [Spi13]. Although it can sometimes be useful, persistent activation can damage a person's health [Lev15]. More detailed information about anxiety can be found in the literature review chapter.

As mentioned previously, anxiety recognition in Virtual Reality is generally done using physiological measures such as SCL and HR and behavioural measures such as respiration and head movement [GGG+19, KSZ21]. Typically, the anxiety within VR is measured by getting the arousal and valence of the user's response; high anxiety is defined as having low valence and high arousal [CXC19]. As such, it is a complicated measure to derive from the user.

The challenge of measuring anxiety accurately in VR comes from multiple reasons. Firstly, it is difficult to accurately define and assess anxiety, as it might differ from person to person [CC67]. Secondly, most established methods of measuring anxiety using physiological sensors require the person assessed to have limited mobility [GGG+19]. Thus, they can prove challenging to use within VR [CXC19].

This chapter aims to develop and investigate a new method for measuring ground truth in anxiety using an adapted version of the well-established emotional Stroop Colour-Word Task.

3.1.1 Anxiety classification and Ground Truth

The ground truth refers to the reference point for the anxiety of the user. Establishing anxiety ground truth is necessary to derive meaning from anxiety classification results. It is standard to use self-report measures such as state-trait anxiety inventory [SGL+83] or subjective units of distress scale [Wol90] for the establishment of ground truth; however, self-report measures come with several biases caused by personal differences and the difficulty of being able to tell one's own level of anxiety [GC89]. Consequently, other methods have become popular, such as using established cognitive tasks, for example, card sort and SCWT [GGG+19].

3.1.2 Stroop Colour-Word Task

Traditional (SCWT) [Str35] is a task that was made for assessing how well a person's brain performs in the existence of conflicting information. The test involves identifying and matching the colour of the words shown where the meaning of the word either matches the colour (congruent or non-interference) or conflicts with the colour (incongruent or interference) [JRJ66]. Based on prior research, most people tend to perform worse in the existence of conflicting information [Mac91]. The time difference in processing conflicting information and non-conflicting information is called the Stroop effect (SE).

Despite being created to assess cognitive ability, researchers discovered that it is an excellent way to measure anxiety and stress [KMY14]. Being such a well-established cognitive task, extensive research has been conducted around it. Today, there are many different versions and applications of the Stroop task. [ST17].

Applications method of Stroop task includes in vivo, computerised and lately, VR [WCL+10, ST17]. Recent research shows that VR applications of the test show similar results compared to the in vivo equivalent [ARE+13, GWM+19].

3.1.3 Emotional Stroop Colour-Word Task

One of the variants of SCWT examines the effect of emotional valence and arousal of word use on reaction time [WMM96]. The so-called emotional Stroop task, sometimes called eStroop, involves words from affective norms for English words (ANEW) [BL99]. Unlike the traditional Stroop task, the emotional Stroop task examines the interference of emotionally loaded words on the processing speed rather than the interference of conflicting information (figure 3.1. This effect is called the emotional Stroop effect (ESE). ESE typically is not as pronounced as SE, where there is a cognitive conflict [KBB96]. However, it effectively creates emotional arousal [MLS+05]. There has yet to be any research on VR validity and interaction of eStroop.

3.1.4 Commercial Devices used for Measuring Anxiety

Based on a review by Gedam et al. in 2021, [GP21] the most popular commercial devices used for the measurement of anxiety include the Shimmer platform for the measurement of GSR, ECG and PPG; Zephyr Bioharness for measurement of ECG; PhysioLab for measurement of EMG, ECG and GSR;

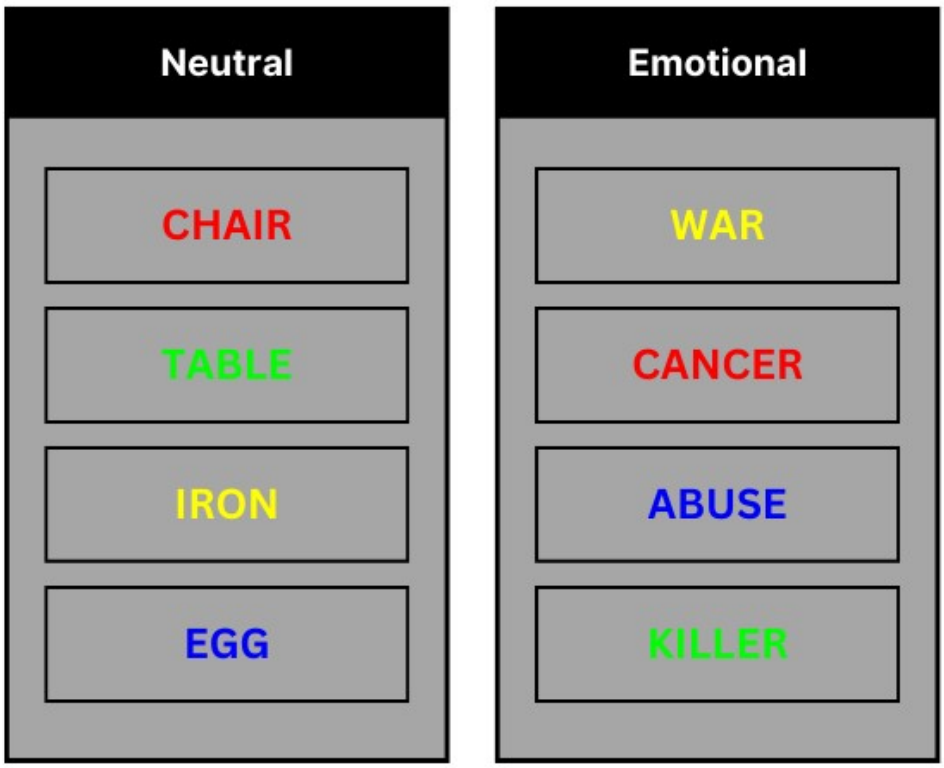


Figure 3.1: Emotional Stroop Task

Polar H7 and H10 chest bands for the measurement of ECG; affective Q-sensor for the measurement of GSR and ACC; Empatica E4 for the measurement of GSR, PPG, SKT and ACC; various Emotiv and Neurosky devices for EEG and Vital Jacket for ECG.

3.1.5 Related Work

Several VR applications show that the performance during VR SCWT applications is similar to performance during in vivo and computerised applications [WCL+10]. As well as testing the effect in a different environment, some of these applications add various elements such as distractions and stressful environments to test different scenarios and how this interacts with the SE [PCAD11].

One of the most well-known examples of VR applications of SCWT is that created by Parsons et al. in 2011 [PCAD11]. In this experience, a similar effect to SE was achieved by experiencing the task in VR [ARE+13]. Furthermore, a

more complex model was developed in the experiment, involving higher and lower threat environments and seeing the interaction between threat level and the SE [PCAD11]. This application is quite different from traditional Stroop applications due to the addition of environmental threats.

Another application of the Stroop Task within VR is by Gradl et al. in 2019 called the Stroop Room [GWM+19]. This application brings a different interaction by changing an entire room's walls to specific colours and letting the users point at them. They have observed a significant increase in heart rate as well as a significant effect on reaction time [GWM+19]. However, it is hard to directly compare the reaction time in this application due to the increased complexity of the interaction.

3.2 Design

This section outlines the application creation process that will deliver the eVRST. The VR application will not only deliver the task and record the responses and reaction times but also the physiological information about the user. Therefore, it is a multi-modal application capable of administering an immersive cognitive task while integrating multiple biosensors to get objective physiological information about the user's state.

This section will outline the process of designing the scene, creating the 3D environment, and developing the application, which includes recording the biological information.

3.2.1 General Design

For this experience, simplicity and robustness are vital. As the goal of the experience is practical and not artistic, minimal attention is given to the details of the modelling or textures. It is meant to test the effect of eStroop in VR so that distractions are avoided, and the focus is on making the experience as barebones as possible. This environment will also be used as baseline detection of anxiety in the VRET framework. The appearance of the scene is modelled based on a waiting room. As the task can potentially induce anxiety for some participants, a relaxation scene with forestry is included in the experience as a fallback environment.

Another critical focus point in this application is the comfort of the user, so

minimal equipment has been used. The absence of clutter will guarantee that the user focuses on the task and possibly impact the feeling of presence in VR positively [MTM21b]. A pipeline of the task flow can be examined in Figure 3.2. The following subsections will detail the decision-making, tools used and application development.

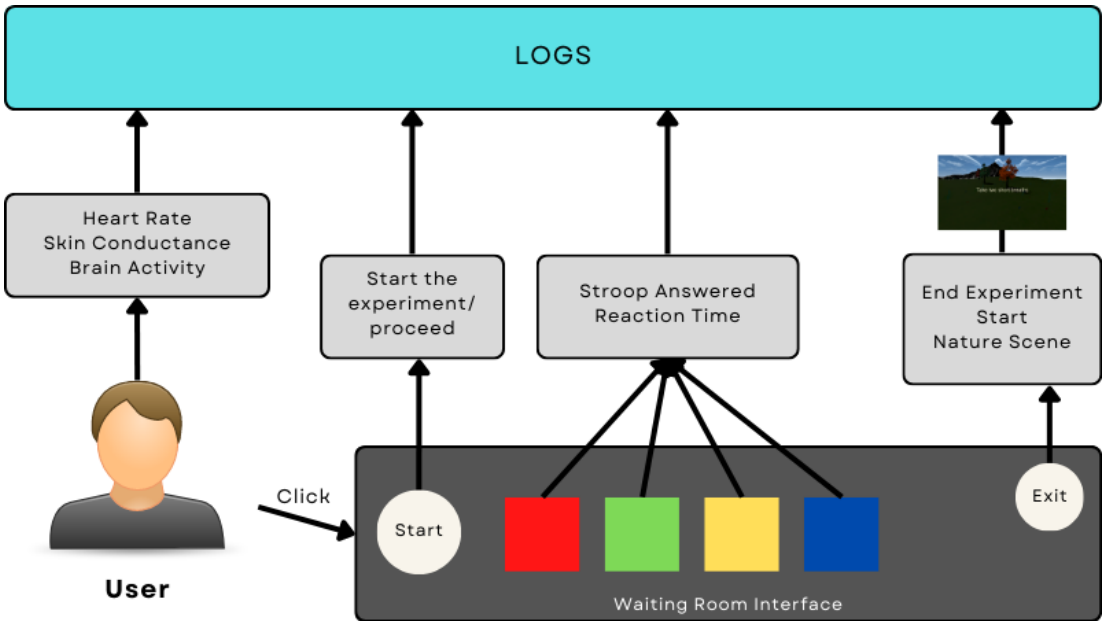


Figure 3.2: eVRST task experience design

3.2.2 3D Engine

The VR application was developed using the Unity 3D engine [Uni22a]. This engine was chosen due to existing resources and ease of development. The project uses the Unity 3D engine to continue work previously developed using the environment [MTM21b]. There are multiple reasons for this choice. First, VR development is very convenient using the SteamVR plugin [Sof19], making it easier to focus on experiment design and human research. Furthermore, the biosignals were processed using Shimmer API [BDG+10], written in C# and conveniently transferred to Unity. Additionally, C# scripting is very efficient and allows developers to implement complex features quickly.

To work on a supported version and not deal with possible unknown problems in newer versions, a long-term support (LTS) version of 2019.4 was chosen. This version is compatible with the SteamVR interaction system, making it easier to develop the application and use it in conjunction with the previous version without full-scale upgrading to a newer version [MTM21b].

3.3 Implementation

This section details the development of the eVRST application.

3.3.1 Stroop Scene

The main scene is modelled to look like a clinic's waiting room to provide many users with a familiar environment (Figure 3.3). The fidelity of the "waiting room" environment is given due consideration. The scene aims for increased fidelity and realism by focusing on correct proportions and plausibility in real life, meaning that, when designed, the elements used in the scene were chosen to represent those who would be seen in real-life scenarios. It is populated by everyday items such as couches, coffee tables, bins, sanitisers and a reception desk. However, no priority is given to the amount of detail and photorealism. Achieving photorealism is not prioritised for two reasons. The first is to avoid the "uncanny valley" effect, as increased similarity to real-life objects can make the user uncomfortable and hyper-aware of small flaws. Second, to measure anxiety based on emotions evoked by a cognitive test, high environment fidelity does not necessarily translate to highly evoked emotions, as the stressors rely on the task itself and the words used in the task. Due to these reasons, photorealism would not be worth the extra development time and a potential decrease in performance and frame rate [SJS14].

The "waiting room" is developed to serve multiple purposes. It is chosen to deliver the cognitive task in this application; however, it also acts as the greeting room for the main VRET framework to be able to record baseline physiological information and decide their course of action. Therefore, it is designed to work for each scenario. There are some subtle changes in the environment applied for each application.

The interaction uses virtual buttons in the environment that can be clicked by moving the controllers on top of them. A hovering platform in front of the user includes a "start/next" button, red, yellow, green and blue colour buttons and an exit button (Figure 3.3). The words are displayed on a television screen placed several meters away from the user.

Most of the main scene's components are created using asset store assets, royalty-free textures, and the Unity package ProBuilder [Uni22b]. ProBuilder allows for the building of environments and objects from scratch within the Unity environment, making it possible to craft the scene entirely within the



Figure 3.3: The waiting room scene, including the interaction platform

Unity Engine. By using ProBuilder, it is possible to avoid using a third-party 3D modelling environment and then importing models into Unity. ProBuilder provides access to basic 3D modelling and mesh manipulation tools such as creating poly shapes, extrusion, UV controls, and vertex editing [Uni22b].

The waiting room is approximately 116 unity square units, equivalent to meter squares. Plenty of space is given to the user to navigate the test easily. The ceiling is three unity units, or meters, high. The room's layout can be seen in Figure 3.4.

In the waiting area, there are several armchairs, a bookshelf, a coffee table, an end table, some coffee cups, some magazines, two televisions, two lamps, two radiators, plants, and a painting (Figure 3.5). All of these resources were free assets downloaded from the Unity Asset store [Uni22a]. The only window in the room is in this area, looking out to a green garden. The outside view changes based on the perspective of the user looking at it, however, this effect was achieved by applying a shader to a 2D image. Creating a 3D garden outside the window is avoided for higher performance.



Figure 3.4: Layout of the Waiting Room

The reception is built using ProBuilders and textured by hand. The textures are created using GNU Image Manipulation Program (GIMP), a highly efficient open-source image editing program [GIM22]. Behind the reception are stationary items such as pencils, pens, a computer, and an office chair sourced from the Unity asset store. For the material of the reception desk, a realistic marble-like appearance was aimed. Therefore, the material source is set to metallic alpha, with 0.25 metallic and 1 smoothness as well as turning on reflections and specular highlights. To correctly reflect the scene, reflection probes are used and baked into the scene's lighting (Figure 3.6).

The Stroop desk is a crucial focus in the room, as the usability of the platform and the screen directly affect the user performance in the experiment. The platform is created using ProBuilder and placed approximately elbow-length for



Figure 3.5: Perspective of a user looking towards the waiting area



Figure 3.6: The reception desk

a person with average height. The function buttons are placed on the platform. The platform contains six buttons: one to start the experiment and proceed between prompts, four colour buttons: red, green, blue, and yellow, and an exit button. One of the most important considerations is the ease of clicking the buttons. Therefore, the buttons are large enough to put the controllers through effortlessly, and the colliders are spaced out far enough to avoid clicking multiple buttons simultaneously by accident. To easily separate the function buttons (start and exit) from the Stroop buttons (red, green, blue, yellow), the functions buttons are circular, and the Stroop buttons are square. When the SUDS feedback option is active, the VR slider replaces the colour buttons. VR slider is made to look like a 3D version of commonly-used sliders on websites and is shaped as a big bright green ball in front of a dull-coloured line (Figure 3.7). A 3D text over the slider with a font size of 72 displays the current anxiety score. More information on SUDS can be found in its respective subsection.



Figure 3.7: The VR slider used to log anxiety

The Stroop screen is a large television screen imported from the asset store with 3D text placed on it. The Stroop text, which displays the information for the Stroop words in the experiment, is placed in the middle of the screen with 128 font size. The script controls the colour of this text. The tooltip text is placed on the bottom of the screen in a font size of 60 and white. The tutorial text, meant to hold the text information during the tutorial, is placed in the middle with size 96.

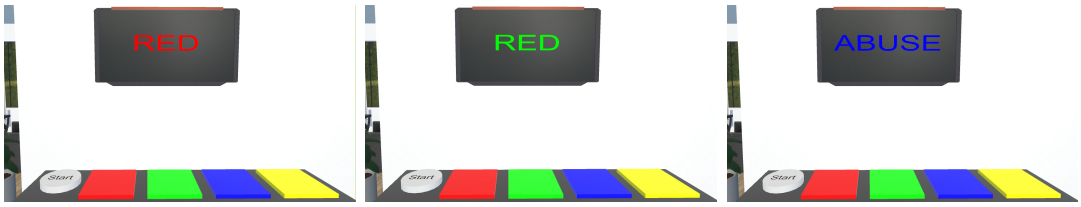


Figure 3.8: The stages of the Stroop task

3.3.1.1 Version History and Changes

After preliminary software testing, some changes were required to increase usability. Most of these changes were to be applied to the Stroop setup. A screenshot of the setup before the changes can be seen in Figure 3.8. Using this scene, it was difficult to spot the changes in colour because of the lack of contrast. Therefore, the white lining was added around the Stroop text. This was added dynamically using scripting, which is detailed in the "task" section. It was also necessary to increase the font size and the television size.

Furthermore, it was difficult for the preliminary testers to see the coloured Stroop buttons that they were clicking. Therefore, the platform was raised and tilted towards the user and scaled upwards to increase usability. Additionally, the exit button was moved further away and made smaller to prevent accidental clicking. The start/next button was elevated higher than the rest to reflect the change of context. Lastly, the user was too close to the screen, so the room was spaced out to be bigger, allowing more distance between the user and the screen. The current state of the Stroop desk and screen can be viewed in Figure 3.4.

The size of the button colliders was set to be narrower and taller to make the response snappier. Changing the collider properties ensures that the buttons would be easy to click independently without touching the other buttons. The participants would not have to put their hands deeply through the button for the click to register; simple contact with the surface would be enough.

3.3.2 Nature Scene

The Nature scene is created to serve as a fallback scene. If the user feels they no longer wish to proceed with the experiment, they can click the exit button to leave the Stroop scene. Once they do so, they are taken into the nature scene until the experimenter removes the equipment. It serves to put the user into a neutral environment without causing any panic. It uses a previously created environment that includes trees, grass, flowers, nature sounds, and slow music

[MTM21b]. It also contains physiological sigh instructions as 3D text and audio. Physiological sigh is a breathing method expected to help people relax quickly. Minimal editing was done to this scene to update the graphics to look more realistic (Figure 3.9).



Figure 3.9: The nature scene

The transition to this scene is started when a user clicks the exit button, and Steam Scene Loader is used to include a fadeToBlack function to prevent the user from potentially observing a semi-loaded scene.

3.3.3 Audio

The audio component of VR is instrumental because it can directly affect the user's experience and presence in the environment [BHvE15]. For this environment, the goal was not to deliver an experience of different worlds but to accurately administer a task. Therefore, the audio in this application is not ambisonic but is bound to objects in the 3D space. There are multiple sources of audio.

As a form of auditory feedback, each button on the interaction platform outputs a subtle clicking noise when touched. The auditory feedback complements the present visual and haptic feedback. The audio source, in this case, is attached to the particular button that is clicked.

There is auditory instructions during the tutorial. Murf AI [AI21] is used to

generate the AI voice-over. Audio is generated using one of their voice models, created to sound similar to a middle-aged British man. Audacity was used to edit the recordings [Tea22].

The nature scene involves some ambient sounds associated with nature, such as the sound of wind blowing and leaves rustling. Instructions in this scene are also generated using Murf AI [AI21]. Audacity [Tea22] was used to edit both the voice-over and ambient sounds in the Nature scene.

3.3.3.1 Version History and Changes

Audio was adjusted based on feedback through preliminary testing as it was necessary to change tutorial steps. The audio's generation and editing methods stayed constant as the audio changed based on the need for a tutorial.

3.3.4 Interaction

The control scheme is developed to be easy to learn and intuitive. It involves no two-dimensional interface features, and all the controls are included in the three-dimensional scene. Interacting with buttons and elements on the screen involves moving the hand (the controller) through the relevant buttons. To accommodate the users' needs, the pacing at the start is set by the user. The user can choose when to start the tutorial and the experiment. An exit button is added to the interface to ensure the user has control over when they can leave the experience.

SteamVR interaction system was used in the experience [Sof19]. The interface elements were all made from scratch, paying close attention to the system's usability. All the buttons in the platform provide audiovisual and haptic cues to provide feedback, ensuring the user is aware of the interaction. The user must move the controller through the button to click the start, colour and exit buttons. During this, the button will light bright yellow, make a clicking noise and vibrate the controller. The vibration of the controller is delivered for 5ms. For the VR slider, the trigger button needs to be held down to make the controller hand into a fist, after which they can move their hand to adjust the place of the slider. The user can see the handle of the slider move as they motion. During the tutorial, the user receives hints on how to use the grab motion. When it is time to record self-reported anxiety, the trigger button flashes light blue on the screen, with the controller changing into a translucent material to make it easier to see. The controller also vibrates until the button is

clicked. This was developed using the SteamVR button hint class, highlighting the button on the controller and providing haptic feedback to draw the user's attention. Once the anxiety is successfully logged, the hints disappear. Users can edit it as often as they like before clicking the "start \next" button to finalise their decision.

3.3.5 Tutorial

The tutorial serves two purposes. The first one is to familiarise the participant with the interaction system and the task by going through every step needed to finish the task successfully. This would also be used to control for learning effects by making sure the users understand the task. The other purpose is to allow the participant to adjust to the VR environment and reduce the novelty effect of first being in an unfamiliar virtual environment. By doing so, the experience aims to minimise the extraneous effect of the environment on the participant's emotional state. This is especially important for users who are not familiar with VR.

When the participant starts the experience, they are placed in front of the interaction platform in the waiting room. The tutorial, for the most part, is self-paced. Before instructions, the participant has time to adjust to the surroundings before the tutorial starts. "Move hand to the "start \next" button to start the tutorial." is written on the screen in front of the user. By allowing the user to choose when to start, their full attention to the tutorial is facilitated. When the user clicks the button and starts the tutorial, the system welcomes the user into the environment by artificially generated speech.

The tutorial is multi-modal; the users can read the text on the screen as well as hear it to enhance accessibility and increase understanding. The experience is briefed in the tutorial. The briefing includes information on the expected duration of the experiment and task instructions. Afterwards, a practical step of the tutorial starts, and the user is invited to make words match until they get four correct answers. This part of the tutorial ensures that the user understands the task. In this part of the tutorial, they receive feedback on whether they match the word correctly or not. Once the user successfully completes this part, they are given ten task-accurate examples of colour matching. By having different steps, including and excluding feedback, the tutorial aims to verify that the user understands the task and will be prepared for the actual task once it starts.

After the no-feedback examples are finished, the participants are asked to input their anxiety level on an SUDS scale. As they need to do a "grab" action for this (in this case, by pulling the trigger on a Vive controller), the controller button for the grab action lights up on the screen, and the controller vibrates. After changing their anxiety level on the slider, the user is able to proceed once again by clicking the "start \next" button. When this step is completed, the tutorial is finished, and the user is invited to start the task by clicking the "start \next" button.

To deliver the tutorial, the "Tutorial Text" object briefed in the graphics section gets dynamically populated using an external JSON file. The JSON file contains two fields: the tutorial text and the hint type. The Tutorial text changes based on the current step and the JSON object's tutorial text field, and the hint text changes based on the hint type field. JSON file format was preferred to make it easy to make quick adjustments as the task steps change. How the JSON file integrates into the tutorial script can be seen in Figure 3.10.

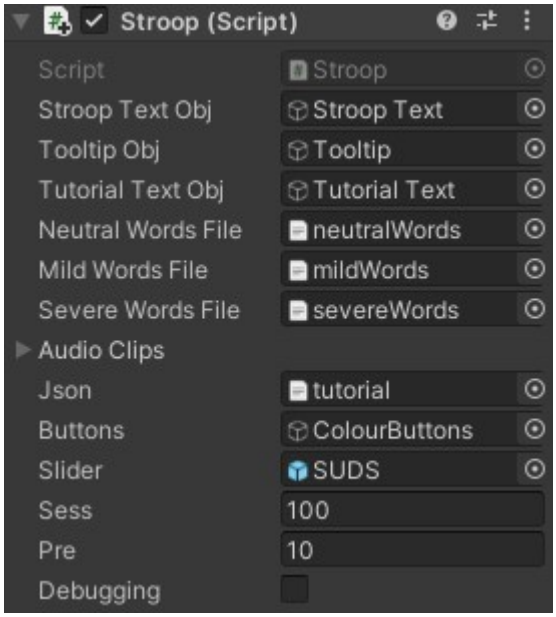


Figure 3.10: Task script options

As this field is dynamically populated; there is no way to adjust the width of the text. To avoid text flowing outside the Stroop screen, a script is added to the item that will wrap the text based on the "line width", which, in this case, is set to the size of the Stroop screen minus some margin. This script works by breaking the long line of tutorial text taken from the tutorial JSON file into smaller lines based on the width given, then updates the graphics that the

participant sees.

Proceeding between the tutorial steps has different requirements based on the sort of instructions. These can be summarised into the following categories: text tutorial, task tutorial, and SUDS tutorial. The condition to proceed during text tutorials is to deliver the text and audio to the participant. The participant does not need to respond to proceed. The task tutorial has two parts: feedback and no feedback. In the no-feedback section, the user must click the Stroop buttons to proceed with the feedback. After receiving their feedback, the next question is displayed. Once four questions are answered correctly, they move on to the no-feedback part. This part is the same as the task display, so they have limited time and proceed regardless of clicking any buttons. If, during this time, they get more than two questions wrong, they repeat the section. After they pass this part of the tutorial, they are taken into their first SUDS entry. To proceed with this, the users need to adjust the VR anxiety slider and click the "start \next" button. This step is the last step of the tutorial, after which the participant is given the option to start the task.

Unity coroutines are used to handle timing and synchronisation in the tutorial. For the tutorial steps that only deliver written and spoken instructions, a countdown is used to navigate the steps. A timer is used for this task. The timer time is adjusted to match the length of the audio instructions being played, with a 1-second padding at the start and end. The tooltip is disabled for this part of the tutorial as nothing is expected from the participant.

An explanation of the task precedes the feedback part of the task and finishes with "here are some examples". This part of the experience calls the Stroop generation task without the time constraint. It waits for user input instead of skipping the question when the time runs out. Once the user answers the question, either "Congrats, correct answer." or "Sorry, wrong answer." is displayed on the screen. The feedback text is accompanied by AI speech. To proceed to the next part of the tutorial, the user is required to answer four questions correctly. To track this information, a counter that increases each time the user gives a correct answer is used. User answers are logged into the .csv log file.

After the feedback task tutorial, the user gets a tutorial text notification that there will be no feedback when the task starts. The tutorial text states, "Here are some examples". The no-feedback part of the task generates ten examples from the Stroop task function. The participant is limited by time constraints

and gets to experience the task more realistically before starting. When it is successfully finished, the user proceeds to the SUDS tutorial.

The SUDS tutorial is also meant to be used as baseline anxiety for self-reporting. The tutorial text changes to "Please log your current anxiety using the slider; 0 = no anxiety; 100; highest anxiety possible.". If the user tries to click the "start \next" button without moving the slider, they are prompted to log their anxiety once again with the tutorial text reading "Please log your anxiety, then click Start \Next". Note that the user can log the same anxiety as the starting level on the slider; however, they need to grab the slider and move to the same location to ensure an answer has been logged.

Once all the steps have been completed, the tutorial text changes to "Tutorial is now complete. You may now begin the task". The tooltip text contains the "start \next" tooltip.

3.3.5.1 Version History and Changes

Previously, manual input by clicking the "start \next" button was required to move between the tutorial steps. Feedback from development testing showed that this made the tutorial tedious and confusing. To prevent this issue, this requirement was removed from the text tutorials and the no-feedback part of the task tutorial. The no-feedback task tutorial was added later to ensure the participant knew what to expect. The tutorial text and Stroop task tutorials were changed multiple times for the changes in the task and updated based on requirements.

3.3.6 Accommodations

As the performance in VRST is dependent on colour vision, colour-changing options were provided for users with impaired colour vision in the editor. Pre-determined colour options from the Unity colour palette were displayed so that colour vision-impaired users could select colours that they could easily distinguish. The option to customise colours was extended to participants with colour vision deficits.

3.3.7 Task

The task is based on the emotional Stroop task, or eStroop, and was modified to match the VR environment. As it can be more challenging to learn a new

task in a VR environment for those who are unfamiliar, utmost care was taken to make it intuitive and enjoyable. The participant chooses when to start the task during the initial period before the tutorial begins and after the successful completion of the tutorial. Once started, the task flows automatically to create a controlled environment and have the same break times for each participant. The exceptions to this are the time to record the anxiety (SUDS) and the use of the exit button. The timing of the task flow can be seen in Figure 3.11.

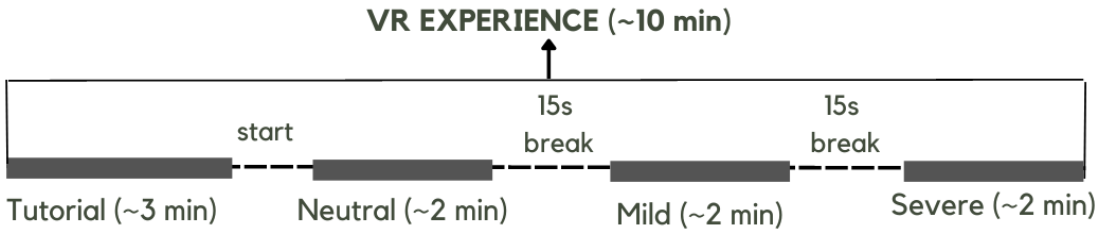


Figure 3.11: eVRST task flow

Unlike the previous versions of eStroop, the eVRST contains two emotional conditions instead of one. So, instead of the traditional "neutral" and "emotional" conditions, this thesis employs "neutral", "mild", and "severe" conditions. While this addition creates complexity and potentially decreases the Stroop effect between each condition, it also allows to identify one additional level of user anxiety.

As such, new criteria for selecting mild words from ANEW were required. Using trial and error, different thresholds of pleasure and arousal were decided. The factors that contributed to creating these criteria were a transition from low arousal and neutral pleasure to high arousal and low pleasure, with a contributing factor of balancing word properties, such as length and count. First, the original ANEW values were standardised and scaled to fit between 0-100. After some experimentation with the parameters, the approved criteria were as follows:

- Neutral: $20 <$ pleasure < 40 AND arousal < 50
- Mild: pleasure < 35 AND $50 <$ arousal < 65
- Severe: pleasure < 35 AND arousal > 65

After the criteria were established, at least 100 words from each group were saved to a text file to be randomly selected during the task. These words can be viewed in Appendix K. Each stage was decided to have 100 words, with 10 example words as a warm-up. The task starts with the warm-up. The number

of words in both the warm-up and main stage are stored in public variables and can be changed directly from the Unity editor. The warm-up is started by tutorial text changing to "Here are ten examples to warm up". After the warm-up, the tutorial text shows "Warm-up is complete. The task starts now". The main stage follows the warm-up with no delay. After 100 task items, users are prompted to enter their current anxiety using SUDS. Before the next stage starts, there is a 15-second break. "Please take these 15 seconds to relax." replaces the tutorial text on the screen. When the 15 seconds have passed, "Break is now over. Get ready for the next stage." replaces the tutorial text. These lines are accompanied by AI speech to focus the user onto the Stroop screen. The code snippet can be viewed in Appendix D Snippet D.1.

An IENumerator Coroutine is used for each stage to handle the warm-up, task, break, and SUDS timing. A numerator is necessary in order to handle time in Unity scripts, such as waiting 2 seconds or an answer before moving between stages. Each item in the task is generated by calling a Stroop generator function (see Snippet D.2 in Appendix D). When an answer is recorded during this time, it automatically switches to the next item.

In the start function of the script, the text files are loaded into the script and shuffled. The Stroop generator function, StroopGen, starts by activating the Stroop text instead of the tutorial text and changes the tooltip to "click the corresponding colour button". If colour buttons are inactive, it deactivates the SUDS slider and activates the colour buttons. The script uses a wordIndex variable to iterate through the items in the array built from the text files: neutralWords, mildWords, and severeWords. The next step calls the colour picker function to decide the colour of the word (see Snippet D.3). The colour picker is called recursively until it picks a colour different from the previous one. Then, a switch statement is called to decide the current stage and correct word array. This stage is passed on to the function as a parameter. Case 1 takes items from neutral words, case 2 from mild words and case 3 from severe words.

Once the colour and the words are picked and displayed, a coroutine is used to manage the time limit. Until 120 ms passes or the user clicks a colour button, the function yield returns null (Snippet D.4). When the user interacts with a colour button, the boolean value "stroopClicked" gets assigned to true. This boolean value is reset back to false in the next call.

3.3.7.1 Version History and Changes

Several significant adjustments were made to the Stroop task after preliminary testing. First of all, the first version of this task mixed emotional and traditional versions of the task for each stage instead of solely focusing on the emotional aspect. This brought unnecessary complexity to the task and made it difficult to differentiate the effects on each stage, as they worked in different ways. As one was focussed on cognitive ability, and the other was focussed on emotional stress, it was decided that the final task would only focus on one type of interference: emotional. Emotional stress was chosen because it carried more transferability to actual use cases than the cognitive load. The change to the emotional stress option came with many subsequent adjustments to the rest of the task. The text files were changed to include varying levels of emotional words. The task script was modified from changing between colour words to only changing the words' colour and picking words from the corresponding text files. The tutorial text and instructions must also be modified to match the current version. The previous version of the task can be seen in Figure D.5.

Previously, the Stroop question did not automatically progress to the next one when answered but waited its entire course of two seconds each time. This was done to ensure the time spent on each stage was equal, however, this was an inadequate solution in terms of user experience. Even with the button tap feedback, the testers felt like their answers were not being registered. An additional solution of adding a visual effect for feedback was considered (Figure 3.12), however, in the end, it was decided that progressing the questions after an answer is registered was the best approach in terms of user experience.

3.3.8 SUDS

The subjective units of distress scale is a scale developed to subjectively measure one's own anxiety from a scale from 0 to 100, with 0 representing no anxiety and 100 representing the highest anxiety possible. A VR version of a UI Slider was developed to log user anxiety. To this end, a line is implemented to represent the range, a green ball is used as the handle for the slider, and 3D text is placed over the slider to show the current level at which it is hovering. No lines or UI elements are used to match the rest of the interaction system; everything is applied in 3D and works similarly to real-life interactions. The user is expected to pull the trigger button on their controller to make their hand into a fist to move the slider around, where the value written over the slider

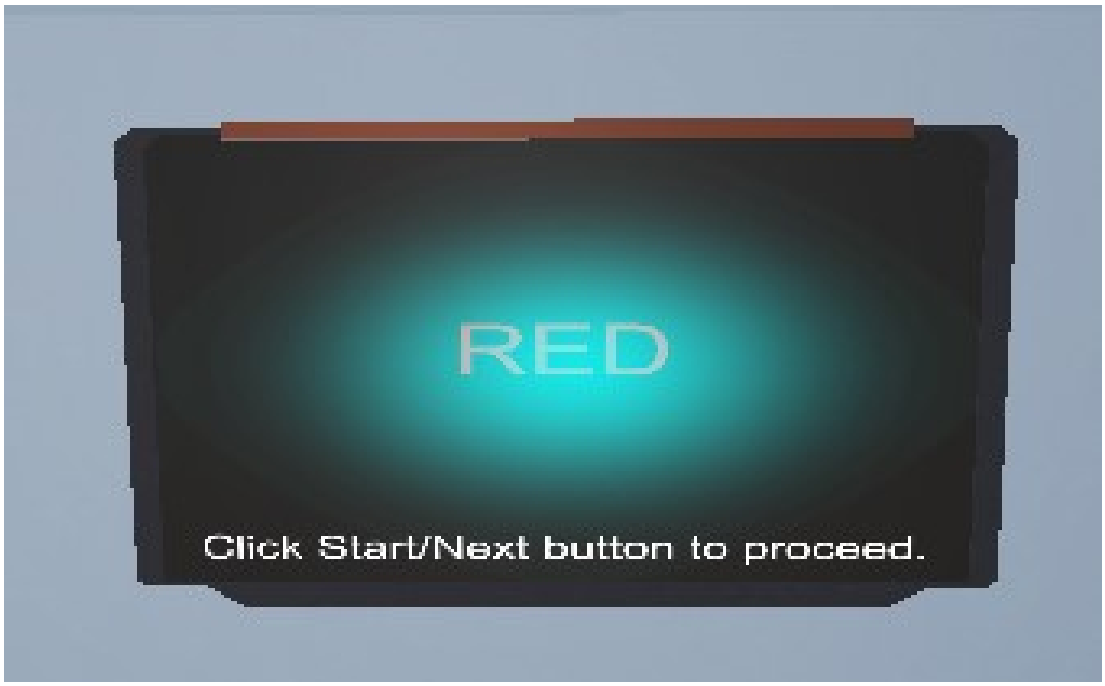


Figure 3.12: Visual feedback for answers

updates in real time.

The script attached to the slider works by getting the relative position of the handler within the range and then mapping the x position of the handle to a range of 0 to 100. The handle only moves when the user has their virtual hand over it in a grabbing motion, and the user stops the grabbing motion to set their anxiety level at the end. If they change their minds, they are able to pick again. Once they are happy with their selection, they can use the "start/next" button to log it. When done so, the anxiety is saved to the .csv file. The Snippet of code that controls the movement of the VR slider can be viewed in Snippet D.6.

3.3.9 Biosignals

Biosignals to be used in the VRET decision-making framework are carefully selected to infer anxiety accurately while maintaining user comfort. The accuracy of anxiety classification is crucial; however, as this thesis focuses on usability and relies on the users wanting to get back into therapy, the accuracy will not be helpful if the user is not encouraged to continue. Therefore, the highest priority is user comfort and mobility, with an emphasis on anxiety classification accuracy. To this end, the physiological information collected was chosen by maximising movement and minimising the weight and devices to be

put on.

3.3.9.1 Metrics and Devices

Pulse rate, EDA, and electrical brain activity are used as metrics. Pulse rate is chosen because of the high correlation it shows with anxiety in the literature [CCF+18]. It is also convenient and comfortable to capture. Skin conductivity, similarly, is selected due to its strong association with arousal levels [LD18]. Electrical brain activity is a more complex measure for VR due to HMD electrical inference and difficulty attaining a good signal without preparation and many channels. However, it is an excellent measure of valence, essential for indicating the direction of the arousal, whether the user is anxious or just excited [YS16]. HRV was not considered for this study because the results were to be used in real-time, and the reliable calculation of HRV can take around five minutes. There has been some research into whether Ultra-Short-Term (UST) heart rate variability measurements can be used as reliably as 5-minute measurements [SMZ20]. Shaffer et al. [SSM16] found that the smallest period that can reliably measure HRV was 60 seconds. A study by Bourdillon et al. in 2017 [BSY+17] similarly found that measurement of HRV with periods below one minute is unreliable.

Skin temperature, head movement, pupil movement, respiratory rate, electromyography, facial expressions, blood pressure, and body movement have also been considered. These measures might have a positive effect on the accuracy of anxiety detection. However, not all of them were suitable for the application for various reasons. Skin temperature, electromyography, and blood pressure were challenging to derive without adding extra devices, which would, in turn, increase the preparation time to use the application and possibly deter people from therapy. They would also increase the total device weight on the user and negatively affect immersion. Despite their possible contribution to anxiety measurement, it was decided that the potential increase in accuracy was not worth the possible negative impact on user experience.

Some metrics, such as facial expressions and respiratory rate, were also unsuitable for use in VR. Facial expressions are complex to derive meaningful information from while wearing an HMD, as only the lips would be visible. Respiratory rate is difficult to derive from nasal devices because many HMDs cover the nose area. There are ways to derive respiratory rate from chest movement; however, these are not lightweight enough solutions for the purpose

of the application. Head movement and body movement were not very useful metrics for the experiment due to the way data was collected. Since eVRST encourages users to move in a particular pattern, this information is not useful or applicable to broader applications.

The devices used for measuring the selected physiological information were decided upon based on accuracy, reliability, and ease of use. Once again, priority was given to user comfort. Multiple devices were considered for each measure.

Skin conductivity is a fundamental metric when attempting to predict anxiety. Thus, the device selection was weighted towards this direction. The options considered for measuring GSR were the Shimmer GSR+ [Sen22] device and the Empatica E4 [Inc24] device. Shimmer GSR+ contains two GSR finger electrodes. Empatica E4, on the other hand, is a wristband that can be attached around the wrist, making it much more comfortable and easier to wear. The preliminary tests aligned with the literature on higher accuracy of signals for the Shimmer GSR+ device. Due to the importance of this signal, the Shimmer GSR+ device was favoured over the Empatica E4 wristband.

For heart rate, the technologies considered were ECG and PPG. Based on prior research, ECG is more reliable and accurate in collecting heart rate information. However, they require multiple electrodes to be attached to the body for optimal signal. Consequently, PPG was chosen in this research due to its compactness and comfort. PPG might produce less accurate results, but it is consistent and valuable. When selecting a device for the PPG, two devices were considered: Shimmer GSR+ and Empatica E4. Empatica E4 provided more comfort and comparable accuracy in heart rate measurements. However, as Shimmer GSR+ provided better results for measuring skin conductance, it was deemed optimal to use this device. As Shimmer GSR+ was already being used for skin conductance, using it to capture PPG ensured the solution was more lightweight and compact. Shimmer GSR+ uses optical pulse sensors to measure PPG. These can be attached either on the earlobe or around the finger. Based on preliminary testing, it was found that the earlobe lead was more reliable.

When it came to electrical brain activity, an EEG was used. Picking the right tool for this task was challenging due to the availability of varying professional and consumer-grade devices. The first decision was how many channels were appropriate to use. More channels provide higher-quality signals; however, they also make the system more cumbersome. It has been decided to use a single channel to simplify the system. When it came to using dry nodes or wet nodes

for the EEG, similarly, the most comfortable option was preferred over higher signal quality and decided on dry nodes. Based on extensive research of literature and available devices, it was decided that the MyndPlay MyndBand [Neu22a] was the best fit for this thesis. This device produces valuable results using a single dry node and a signal channel. MyndPlay MyndBand features three sensors, a headband and a Bluetooth unit. The headband was removed, and the Bluetooth unit and sensors were attached directly to the HTC Vive unit to make it more comfortable for the users. This made it easier to use and avoid problems such as the headband slipping with HMD movement.

3.3.9.2 Data Acquisition and Real-time Processing

As described in the previous subsection, Shimmer GSR+ was used for PPG and EDA and MyndPlay MyndBand was used for EEG. PPG has a training period of ten seconds, whereas the other measures can be used earlier. The EEG has a refresh rate of 512 Hz, where data from PPG and EDA streams every second. Multiple methods of data acquisition were considered. Firstly, the raw data from EEG was acquired using OpenVibe and separated into different wavelength bands using filters. Then, it was synced with the information from Shimmer GSR+ using LabStreamingLayer (LSL). This method was efficient, however, it was laborious and lacked validation. In favour of time and scientific reliability, the API provided by NeuroSky, called ThinkGear Controller was employed.

ThinkGear API processes the raw EEG data and outputs them every second in a format that is only meaningful when compared to itself and has no unit. It should be noted that while it is useful for recording changes intrapersonally, it is meaningless interpersonally. The wavebands are recorded as follows: delta from 0.5 to 2.75 Hz, theta from 3.5 to 6.75 Hz, low alpha from 7.5 to 9.25 Hz, high alpha from 10 to 11.75 Hz, low beta from 13 to 16.75 Hz, high beta from 18 to 29.75 Hz, low gamma from 31 to 39.75 Hz, and mid-gamma from 41 to 49.75 Hz. As the data outputs every second, it smoothly syncs with EDA and PPG data.

Heart rate is acquired and processed using the Shimmer C# API. Raw PPG is run through a Blackman Window-Sync filter to remove frequencies over 5 Hz and below 0.5 Hz. The filtered PPG signal is then used to calculate the heart rate using the PPGtoHeartRate function. If the signal is not clear enough to be processed, it outputs -1 as the result of the calculation. In this case, the last valid heart rate calculation within five seconds is used.

EDA is also acquired through the same Shimmer C# API. This outputs the

skin resistance in kOhms, which is converted to skin conductance in real-time, measured by microSiemens.

The biosignals, user answers to questions, and events are all recorded into an external .csv file and timestamped accordingly.

3.4 eVRST Testing

The software was tested using two hardware sets, and it was able to maintain a stable 100 frames per second on each set. The first used an HTC Vive HMD, GeForce GTX Titan X graphics card (GPU), and Intel i7-5820k processor. The second used an Oculus Rift S HMD, RTX 3070 GPU, and Ryzen 5 3600X processor. The controls were responsive and worked as intended in both cases.

The application was tested on two testers for feedback during the development. An open-ended interview was used to record their feedback to improve the application. The interview included questions about the comfort of the hardware, intuitiveness of the tutorial, interface usability, and general feedback. In terms of comfort, both testers said they were aware of the sensors at the beginning but got accustomed to them as the experience went on. One of the testers, due to their height, said they had difficulty looking at both the buttons and the screen simultaneously. Based on feedback, the distance between the screen and the user and the height of the buttons were adjusted. Another tester mentioned that the "start \next" button was too close to the colour buttons and should be separated. Following this feedback, the height of the "start \next" button was adjusted.

3.5 Discussion

In order to achieve research goals, a few priorities were laid out for the development of the eVRST environment. Firstly, it must identify three distinct conditions that can later be used to classify three anxiety levels. Secondly, this effect must be verifiable using both objective and subjective measures. However, it is possible to interrupt the flow of the experiment by adding long questionnaires and questionnaires that require the participant to leave the environment. Therefore, subjective and objective measures needed to be smoothly integrated into the VR environment. This was achieved by logging objective and subjective measures, time, and variables in a log file through

Unity and third-party APIs.

In the design of this application, three components were crucial. First of these was the accessibility of the application. As it can be challenging for some individuals to take advantage of VR therapy due to factors such as age, education level, and visual deficits, utmost attention was given to eliminate and minimise the risks of these factors [HHK18].

An essential factor of accessibility is simulator sickness. Simulator sickness can impair engagement, performance, and enjoyment in VR, and it affects a significant portion of the population [SSB+20]. This is especially pronounced for users who are not accustomed to VR. In order to combat simulator sickness, which affects user experience negatively, the following steps were taken. The VR design was simplified to adapt a cartoonish look as opposed to photorealism. The Field of View (FoV) was set to industry standard using SteamVR camera. Natural movement, as opposed to using controllers, was employed to avoid discrepancies between visual and body movements [CKY20]. The experience duration was kept at a lower than thirty minutes. The frame rate was set at a standardised level (100 frames per second), which is stable, and the participants will not experience frame drops during the experience.

To accommodate participants who are less experienced with VR and those within a larger age range, multiple strategies for button collision feedback and controller hints were incorporated into the application. Haptic, auditory, and visual feedback was utilised to inform participants of button collisions [IKJM+19]. The tutorial was thorough and self-directed to allow participants to fully understand the task as well as get used to navigating a virtual environment. Tutorial instructions were both displayed in text and read out using audio to accommodate participants who might have difficulty reading or hearing.

The second crucial component is the safety and comfort of the participants during the experiment. Participants were guided to an optimal location in the laboratory with plenty of space to move around and no obstructions in the immediate surroundings. The experience was set up in a way that would not require extensive movement. Additionally, for comfort, the sensors were carefully selected and placed. The wrist-based sensor, Shimmer, was placed on the non-dominant hand to increase the quality of experience for participants and decrease noise. The Neurosky sensor was attached directly to the HMD to avoid extra irritation.

The third crucial component was ensuring user engagement. Some strategies were adapted to optimise the participant's engagement with the task. Firstly, the users were allowed to start at their own pace. Self-direction promotes ownership and a sense of agency [KGS12]. It also helps minimise the initial anxiety by ensuring that the participant is ready to face the challenge. A smooth onboarding experience was promoted, utilising a thorough tutorial and initial instructions. The length of the experience was optimised to promote a flow state [CAN05].

The design of the eVRST experience has some similarities and dissimilarities with previous work on VR Stroop adaptations. For instance, instead of using a multi-item Stroop presentation, a single-item Stroop presentation was utilised in order to examine individual reaction times to different words [PB18]. Unlike some previous work, this environment did not include added distractors [PCAD11]. The Stroop examples were presented using a virtual TV in the environment for its similarity to computerised applications. Complex interactives such as coloured walls that participants need to move around to select were avoided for better comparison to previous work [GWM+19].

3.5.1 Limitations

While accessibility was prioritised in the design of this environment, there are some understudied aspects of using cognitive tasks for anxiety classification. Some research suggests that people with higher cognitive function might not be as anxious while completing cognitive tasks. Additionally, the Stroop effect might or might not be diminished for individuals affected by learning disabilities such as dyslexia. To control for these effects, a demographic questionnaire was added to the pre-questionnaire, and the results were monitored in light of dyslexia, colour blindness, and education level.

3.6 Summary

This chapter showcases the design, development, and testing of a novel virtual reality application for the well-established eStroop task. This environment aims to provide a new, generalisable, less biased way of deciding the ground truth for anxiety classification. In order to be used as a ground truth, the system first needs to be validated. Another function of this application is a novel way of applying the well-established eStroop cognitive task. After validating the

eStroop task and a comparison to previous studies, the platform can be used as a convenient way of applying the task within VR. Although in vivo and computerised applications are sufficient for general applications of the task, this allows integration of the study into VR applications, furthering its reach, and providing a way of including it in VR research and entertainment.

In the following chapter, this environment is evaluated based on usability and analysed to examine the differences in self-reported anxiety, physiological anxiety, and performance across different conditions: neutral, mildly emotional, and severely emotional.

The design of this environment is a big step towards the thesis's primary research goal: to provide real-time, objective feedback on anxiety. Once validated, this environment will not only lay the foundation for classifying between varying levels of anxiety but it will also be used as a means of establishing the baseline for personalisation and normalisation during exposure therapy. Therefore, this chapter also contributes to answering sub-research question 1: Can ground truth in anxiety be made more bias-free and accessible via the use of cognitive tasks?

The next chapter will build upon the current chapter to validate the eVRST application in terms of its similarity to in-vivo applications. The next chapter will also examine the three conditions outlined in the development (neutral, mildly emotional, and severely emotional) in terms of participant arousal, performance, and self-reported anxiety.

Chapter 4

eVRST: Emotional Virtual Reality Stroop Task Validation

This chapter explains the evaluation and validation of the e-VRST task described in the previous chapter. It details the steps of experiment design, materials, participants, data collection, and analysis. This experiment aims to prove that 1) the eVRST system produces similar results to previous in vivo applications of e-Stroop and 2) the three conditions of the eVRST are significantly different in terms of self-reported anxiety, skin conductivity, and performance. If the results align with previous work, the information collected will be used as a basis for machine learning analysis. This chapter extends the publication entitled "Emotional Virtual Reality Stroop Task: Pilot Design", published in Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology (VRST '21) in 2021 [MTM21a] and "Emotional Virtual Reality Stroop Task" published in the IEEE International Conference on Pervasive Computing and Communications in 2023 [MTM23b].

4.1 Background

This study examines the differences in means of electrodermal activity, self-report anxiety and reaction time in different stages of the developed task. By examining these differences, the aim is to find out whether the eVRST application is suitable for differentiating between different levels of anxiety. For physiological measures, SCL is used as a marker for high arousal levels. Higher arousal is interpreted as negative based on self-reported anxiety levels.

For the eStroop task, a blocked design was used to focus on the general effect over a time period [BHWH+16]. Additionally, a fixed order instead of a mixed order was used because mixed order does not allow the recording anxiety changes due to sustained effects of the emotional words [BHWH+16].

4.1.1 Hypothesis & Predictions

The neutral, mildly emotional, and severely emotional conditions should be significantly different in order to use machine learning to distinguish between them. Therefore; the hypotheses are:

1. There will be a significant difference in the mean levels of GSR between neutral, mildly emotional, and severely emotional conditions. It is predicted that the mean GSR will be highest during the severe emotional condition and lowest during the neutral condition.
2. There will be a significant difference in the mean SUDS scores between the neutral, mildly emotional, and severely emotional conditions. It is predicted that the mean SUDS score will be highest during the severely emotional condition and lowest during the neutral condition.

4.1.2 Self-reported Anxiety

In the literature, many studies use separate methods for self-reporting anxiety. All of these methods aim to measure similar constructs but use different conducts. There are multiple factors to consider when choosing a method of self-reporting. There is the validity and reliability of the measures, the cost, comparability and length [SCCC14]. Due to the different scoring systems and methods between measures, it might not be easy to compare them directly. The most used options are STAI, Mood and Anxiety Symptom Questionnaire (MASQ), Generalized Anxiety Disorder Scale (GAD-7), Positive and Negative Affect Schedule (PANAS), and SUDS.

STAI is a 40-item scale rated on a Likert scale from 1 to 4 on how they feel [SGL+83]. It is the most well-known measure for trait anxiety [KO20] and has a Cronbach's alpha reliability score of .85. It has many modified versions, including STAI-S, which measures current feelings and STAI-T, which measures general feelings. There are also shorter versions, such as the STAI-Y1, which includes only six items [MB92].

MASQ is a 90-item Likert-scale with subtypes of anxious, depressive and mixed

symptoms [WC91]. The Cronbach's alpha for the 38-question scale was measured at .94 [KSGW15]. There are many subscales and modified versions that are used for those who do not intend to include the full 90 items. MASQ-D30 is a 30-item shortened version developed by Wardenaar et al. [WvVG+10]. MASQ-GA is an 11-item General Distress-Anxious Symptoms scale [SCCC14].

GAD-7 is developed as a diagnostic tool for generalised anxiety disorder. However, it is often used as a measure of general anxiety in studies [SKWL06]. The internal consistency of this questionnaire, measured by Cronbach's alpha, is .92. The oftenness of symptoms in the last two weeks is rated from 1 (not at all) to 4 (nearly every day).

PANAS is a 20-item scale that measures positive and negative emotions by selecting how much of a certain emotion word the participant felt in the last seven days on a scale from 1 (not at all) to 5 (extremely) [WCT88]. It shows high internal consistency with a Cronbach's alpha score of 0.86 to 0.90 for the positive scale and 0.84 to 0.87 for the negative scale. PANAS-X is the extended version with 60 items and distinct emotion categories such as fear, sadness, guilt, hostility, shyness, fatigue, joviality, self-assurance, attentiveness, serenity and surprise [WC94]. The words that are related to fear include afraid, scared, frightened, nervous, jittery, and shaky.

4.1.3 Related Work

There are several VR applications that show SCWT performs similarly in VR as it does in vivo [WCL+10]. As well as testing the effect in a different environment, some of these applications add in elements such as distractions and stressful environments to test different scenarios and how this interacts with the SE [PCAD11].

One of the most well-known examples of VR applications of SCWT is that created by Parsons et al. in 2011 [PCAD11]. In this experience, a similar effect to SE was achieved by experiencing the task in VR [ARE+13]. Furthermore, a more complex model was developed in the experiment, involving higher and lower threat environments and seeing the interaction between threat level and the Stroop effect [PCAD11]. This application is quite different from traditional Stroop applications due to the addition of environmental threats.

Another application of the Stroop Task within VR is by Gradl et al. in 2019

called the Stroop Room [GWM+19]. This application brings a different interaction by changing an entire room's walls to specific colours and letting the users point at them. They have observed a significant increase in heart rate as well as a significant effect on reaction time [GWM+19]. However, it is hard to directly compare the reaction time in this application due to the increased complexity of the interaction.

Some examples of previous studies that investigate the Stroop effect in different formats can be found in Table 4.1.

Table 4.1: Previous work on Stroop and eStroop tasks

Reference	VR?	Stroop Type	SCL incr. ($\mu\%S$)	SE (ms)
Renaud et al. [RB97]	NO	Stroop	NS[a]	S[b]
Malhi et al. [MLS+05]	NO	eStroop	26.6	1.4
Strauss et al. [SAJC05]	NO	eStroop	n/a	-2
Eilola et al. [EH11]	NO	eStroop	10	38.5
Blanchette et al. [BR13]	NO	eStroop	5	S[b]
Gradl et al. [GWM+19]	YES	Stroop	12	910

[a]Not significant (rates unreported), [b]Significant (rates unreported)

4.2 Experiment Design & Methods

4.2.1 Participants

Twenty-nine volunteers participated in the study. The age distribution of the participants is as follows: eleven participants between the ages eighteen and twenty-four (38%), fifteen participants between the ages twenty-five and thirty-four (51%), two participants between the ages forty-five and fifty-five (7%) and one participant older than 55 (3%). These figures can be observed in Figure 4.1. Eleven participants were female (38%), and eighteen participants were male. There were seven participants who never used VR (24%), fifteen participants who used it a couple of times (51%), five participants who had one to two years of experience in VR (17%) and two participants with two to five years of experience (7%) (Figure 4.2). Seven participants stated they never experienced motion sickness (24%), fourteen participants experienced it rarely (48%) and eight participants sometimes (28%). Four participants were diagnosed with colour blindness (14%), and three participants were diagnosed with dyslexia (10%).

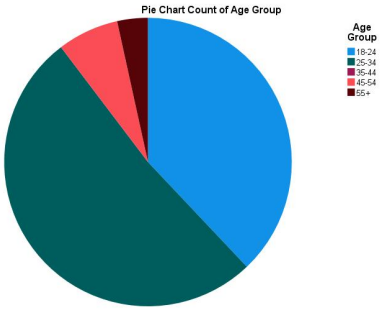


Figure 4.1: Distribution of age groups among participants

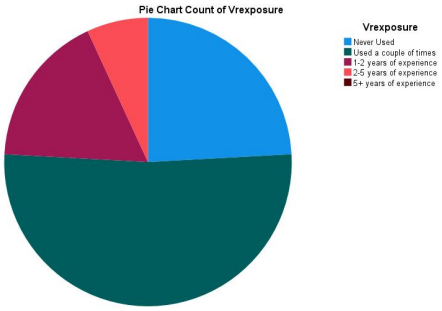


Figure 4.2: Distribution of prior VR exposure among participants

4.2.2 Materials

Two questionnaires were administered: one before the experiment and one after. The first questionnaire included demographics and a short version of the State-Trait Anxiety Inventory (STAI-Y1) scale. State-Trait Anxiety Inventory Y1 was developed by Spielberger et al. [SGL+83] and modified to be shorter by Morteau and Bekker [MB92]. In the VR experience, subjective units of distress scale (SUDS) was used to log anxiety as a number between 0-100. The second questionnaire included subjective classification of each stage in eVRST and Likert-style subjective questions. The consent form, pre-questionnaire and post-questionnaire can be viewed in appendices E, F and G, respectively.

A Shimmer3 GSR+ device was used to measure SCL, and an optical pulse ear-clip add-on was used to measure PPG, which was then used to estimate heart rate [Sen22]. Shimmer API was used to process these signals [BDG+10]. MyndPlay MyndBand was used to collect electrical brain activity [Neu22a]. NeuroSky API was used to process raw signals [Sał14]. A participant wearing all of the equipment can be viewed in figure 4.3.

The application was developed and tested using the HTC Vive HMD and a personal computer with GeForce GTX Titan X as the graphics card and an Intel i7-5820k processor. The audio was delivered via the HTC Vive earphones.

4.2.3 Procedure

Participants first were given an information sheet and consent sheet. They were given the first questionnaire once they were satisfied with the information and signed the consent form. These questionnaires were assigned a unique ID to match their physiological information and the second questionnaire. After this point, they were moved to the ”play area” for the VR part of the experiment.



Figure 4.3: A participant with the equipment attached

The experimenter placed the Shimmer GSR+ device on the left wrist, connected the electrodes to the middle and ring fingers, and connected the optical pulse to the earlobe. Afterwards, the HMD was placed on their head until they felt comfortable and had no blur in their vision. Then, the earphones were attached, and the controllers were handed to them. They were given instructions on how to leave the VR environment in the event that they felt uncomfortable during the experiment. Then they started the tutorial when they felt ready. They experienced the tutorial, the three stages, and the breaks between them. After they had completed the experiment, the experimenter removed the equipment from them. Once they felt ready, they were given the second questionnaire. After it was all done, they had the opportunity to ask the experimenter questions.

For participants who reported being colour-blind, the colours were adjusted to ensure a quick distinction. This was done by first showing these participants the current selected colours and then allowing them to adjust the colour options until they were satisfactory.

4.2.4 Preparation of the data

As outlined in the previous chapter, the following information was logged in .csv files per each participant: timestamp; electrical brain activity bands corresponding with low alpha, high alpha, low beta, high beta, delta, theta, low gamma, high gamma; skin conductivity level; heart rate; current stage of eStroop and current word on the screen. This information was sampled in a 1-second timeframe. Additionally, event-based responses were recorded on these CSV files when the participant answered a question. For the Stroop answers, this included the timestamp, reaction time and whether their answer was correct. For self-report measures of anxiety, the anxiety was recorded.

First, this information had to be laid out in a format suitable for mean comparison between the stages, as per the main aim of the experiment. IBM SPSS was used for data entry and cleaning. First, the variables to be used, their data types, labels and values were created. The names, types and labels of these variables can be found in Appendix H. Nominal measure refers to a variable with separate groups, ordinal measure refers to a variable with relative measures from each other but not to scale, and scale measure refers to scalable data.

After the variables were logged into the SPSS, they were populated. First of all, the data that did not need processing were directly entered into SPSS from the forms. These variables were demographic information including ParticipantCode, Agegroup, Gender, VRExposure, MotionSickness, ColourBlindness and Dyslexia; questionnaire/psychological scale information such as STAI, SelfVR1, SelfVR2, SelfVR3, Self1, Self2, Self3, and Q1-12; accuracy of Stroop answers which was calculated within Unity and logged onto .csv: Acc1, Acc2, Acc3; and finally, Comments. Comments were categorised into common sentiments before putting it in to make it easier to report.

For ease of use, convenience and automation purposes, the scale variables were calculated using Python and then logged into SPSS. For these calculations, all the .csv files were collected into one dataset, with an extra field for participant code added. For the reaction time, instructions for ESE calculation from Ben-Haim et al. published in 2016 [BHWH+16] were used. First, the false responses were eliminated using substring matching. Then, outlier removal was applied, removing reaction times 2.5 standard deviations from the mean. The resulting reaction times were then separated into stages and grouped by participant code. Means were taken per participant and stage and were entered into SPSS, occupying RT1, RT2, and RT3 variables. OverallRT was also

calculated in Python by removing the stage separation. Overall ESE was calculated in Python using "mean reaction time in stage 3 - mean reaction time in stage 1".

A similar procedure was followed for SCL values. Pandas and NumPy were used on the aforementioned data frame, but this time, SCL values three standard deviations away from the mean were removed based on the means and standard deviations of each participant, as opposed to the entire data frame. The reason for doing this per participant was the possible interpersonal differences in skin activity among the participants. The resulting mean values were entered per participant and stage, occupying GSR1, GSR2 and GSR3 variables.

SelfDiff was calculated in SPSS using "SelfVR3 - SelfVR1". ESE per user was calculated in SPSS by "RT3 - RT1". SCLpercent was calculated within SPSS, using "(GSR3/GSR1) - 1".

4.2.5 Performance-related Measures

This study uses ESE, calculated by getting the mean reaction times of correct responses and removing outliers, as described previously. However, when it comes to questions where accuracy and reaction time are involved, there are multiple ways to analyse performance. The best methods for this were analysed and compared before a decision was made. These methods will be summarised in this section.

A common approach in early cognitive evaluations would be to separately analyse reaction time and accuracy [BB11]. This might, however, make the results more challenging to compare by having to manually combine two separate results, especially if they are conflicting.

One method proposed by Townsend & Ashby in 1983 [TA+83] combines these two metrics by dividing mean reaction time by per cent of correct answers. This method can be useful for certain distributions of data, however, it is not advised to be used when the error rate is high [BB11, TA+83]. As this study includes some participants with error rates above 10%, it was decided that this method is not appropriate.

Using the Analysis of Variance (ANOVA) formal test to compare variances of reaction times is also a common method for analysing this type of data. Whelan (2008) [Whe08], however, warns that ANOVA tests might be unsuitable due to their power possibly being too low to detect differences in RT. Whelan [Whe08]

recommends investigating the entire distribution for insights.

Diffusion models are also used to combine metrics such as variance in reaction time, reaction time and response accuracy. One of the most common diffusion models is the one proposed by Wagenmakers et al. [WVDMG07]. To be able to apply this diffusion, the model assumes that the data is right-skewed (increasingly more so as the task gets harder), that the relative speeds of correct and wrong responses are similar, and the starting point is unbiased [WVDMG07]. The data is misspecified for these assumptions, as the reaction times are not right-skewed. Skewness of the reaction time per stages 1, 2 and 3, in order, are: .079, .953, .663 using formal tests in SPSS.

Due to the unsuitability of the common combining options and the difficulty of interpreting two different metrics, the ESE technique was adopted from the research of Ben-Haim et al. [BHWH+16].

4.3 Descriptive Analysis

This section describes the methods for preparing and processing the data collected from the eStroop experiment, the analysis of said data using appropriate methods, extra considerations and analysis that can be insightful for future research and the interpretation of the analysis carried out. Analysis was carried out using IBM SPSS, Python, and Python modules, such as Pandas, to use data frames and Numpy for math calculations.

4.3.1 Descriptive Statistics

4.3.1.1 Recorded Variables

Variables were explored to find information about the ranges, means, medians and variances of the variables. The STAI scores ranged between 0-11, the median score was 1, the mean was 2.38, and the standard deviation was .51. Histogram can be viewed from Figure 4.4. One participant was marked as an outlier with a score of 11. Removing this outlier was unnecessary, so it was kept in the dataset. The STAI distribution is skewed to the right. Logarithmic transformation was applied to the data to derive meaning from the STAI results. The new distribution of log values can be viewed in 4.5. It can be interpreted that the distribution was normalised. However, due to the number of 0 values in the STAI scores of the sample, logarithmic transformation turned

ten values into missing values. Since the method is not applicable to a large portion of the data, it was determined to be an inappropriate transformation. Instead, square root transformation was applied to see if it would be a better-suited normalisation. The distribution of the square root values can be viewed in Figure 4.6. It can be seen in the histogram that the distribution still doesn't follow a normal curve. It was deduced that the STAI distributions were not fit to be transformed to follow a normal curve without significantly changing the data.

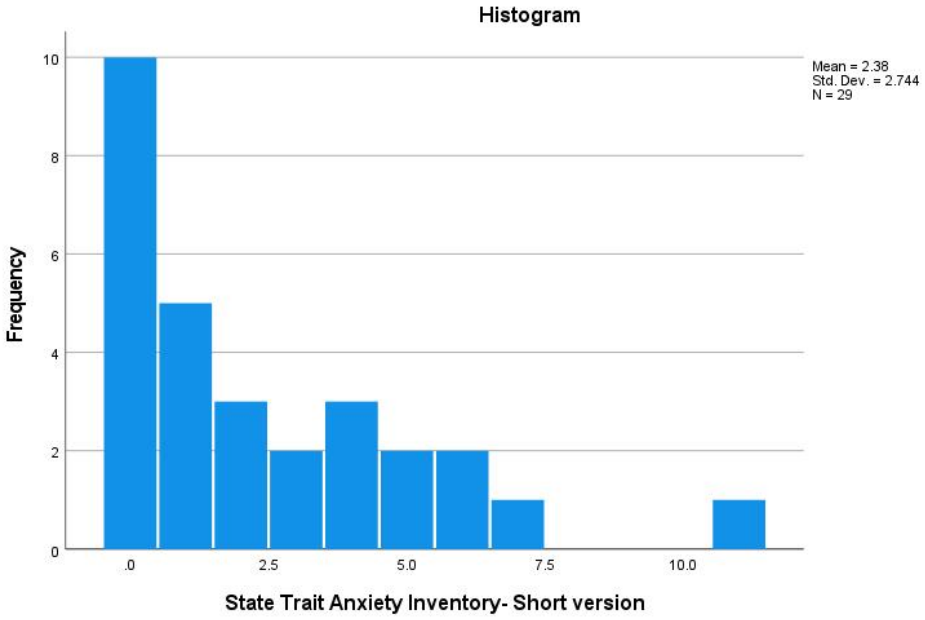


Figure 4.4: Distribution of STAI scores among participants

The mean SCL, RT and SUDS in each stage of the task can be viewed in Table 4.2. From observing these values, it can be seen that there is an increase in SCL values as stages progress. Reaction time has decreased, which was unexpected. SUDS values have slightly increased. Formal tests need to be conducted to further interpret the mean changes.

There were also numerous subjective questions studying the usability of the system on a Likert scale of 1-5. The responses to these questions can be observed in Table 4.3.

4.3.1.2 Computed Variables

Minimum, maximum, mean and standard deviation values of the variables SUDS increase, percentage SCL increase, overall reaction time and reaction

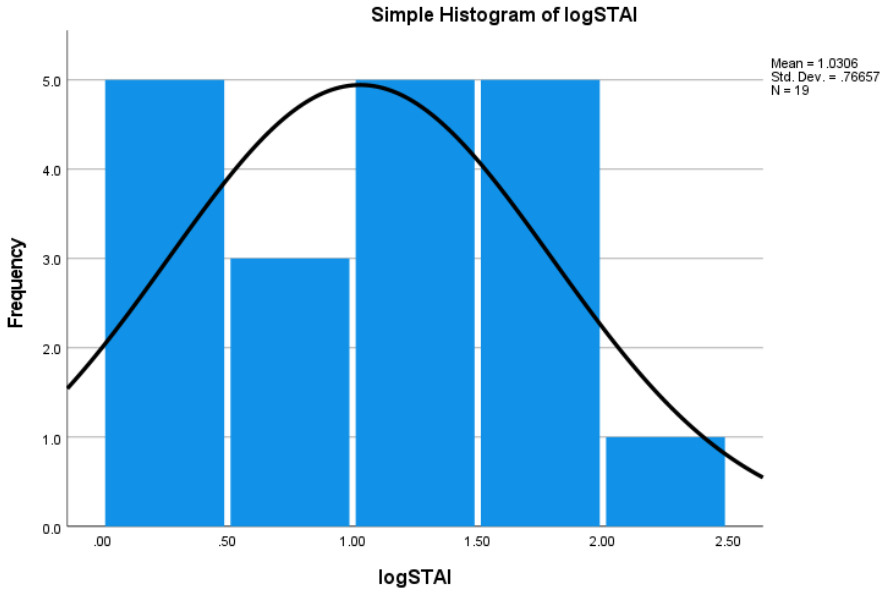


Figure 4.5: Distribution of log-transformed STAI scores

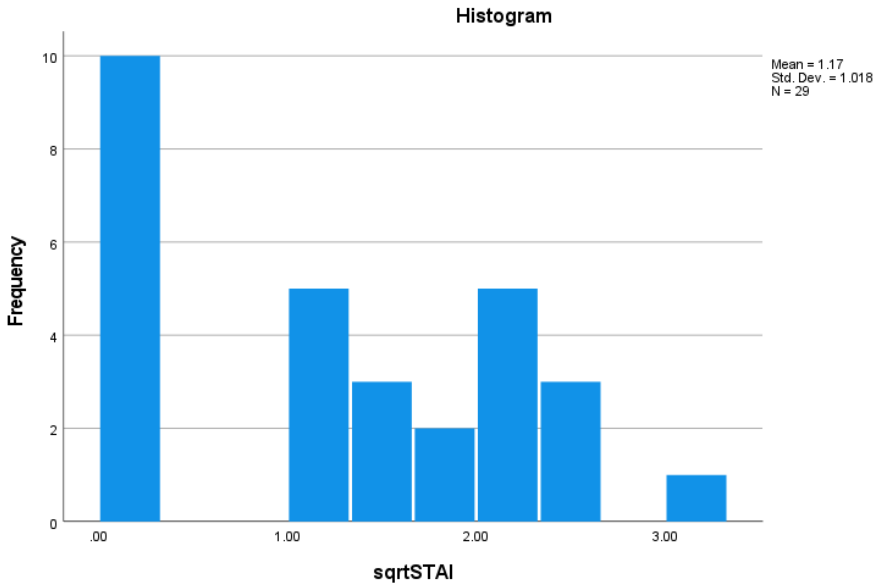


Figure 4.6: Distribution of square root transformed STAI scores

Table 4.2: Descriptive statistics of SCL, RT and SUDS between each stage

	Mean	SD	N
Average GSR in stage 1	3.00	2.35	29
Average GSR in stage 2	3.58	2.99	29
Average GSR in stage 3	4.00	3.32	29
Reaction time in stage 1	986.64	150.29	29
Reaction time in stage 2	983.58	137.12	29
Reaction time in stage 3	956.87	114.99	29
SUDS after stage 1	19.83	17.63	29
SUDS after stage 2	21.03	16.89	29
SUDS after stage 3	23.83	19.13	29

Table 4.3: Questionnaire responses

Question	Mean	N	SD
I found the equipment worn during the experiment cumbersome and uncomfortable.	2.72	29	.960
I was able to understand the instructions clearly.	4.59	29	1.053
I was conscious of the devices on me at all times.	3.17	29	1.104
The amount of equipment I was wearing reduced my immersion in the virtual world.	2.45	29	.870
I felt bored during the experiment.	2.07	29	.961
I was comfortable with the devices on me.	3.62	29	.820
I felt motion sick during the experiment.	1.21	28	.418
I found the platform with the buttons easy to use.	4.10	29	.860
I found the task confusing.	1.48	29	.634
The experiment had me engaged at all times.	4.14	28	.756
I wish the tutorial was more thorough.	1.72	29	.649
I found it difficult to understand and carry on the task given to me.	1.28	29	.455

time increase can be viewed in Table 4.4.

It can be observed that SUDS change has ranged between decreasing by 18 to increasing by 22. On average, the SUDS increased by 4, with a variation of 10. Upon examining the boxplot (Figure 4.8, two participants were identified as outliers. Both outliers stated they were much less anxious in stage 3 than in stage 1. Upon examination, neither of these participants is an outlier in terms of ESE or SCL % increase. A histogram of SUDS increase was drawn to examine the distribution of the data (Figure 4.7). The black line shows the normal distribution. It can be deduced that the data distribution is even

Table 4.4: Descriptive statistics of computed information

	Min.	Max.	Mean	SD
SUDS increase	-18.00	22.00	4.00	9.91
% increase in SCL	-.49	1.00	.31	.29
Overall Reaction Time	790.61	1257.42	975.70	124.92
Reaction time increase	-267.32	172.97	-29.78	95.46

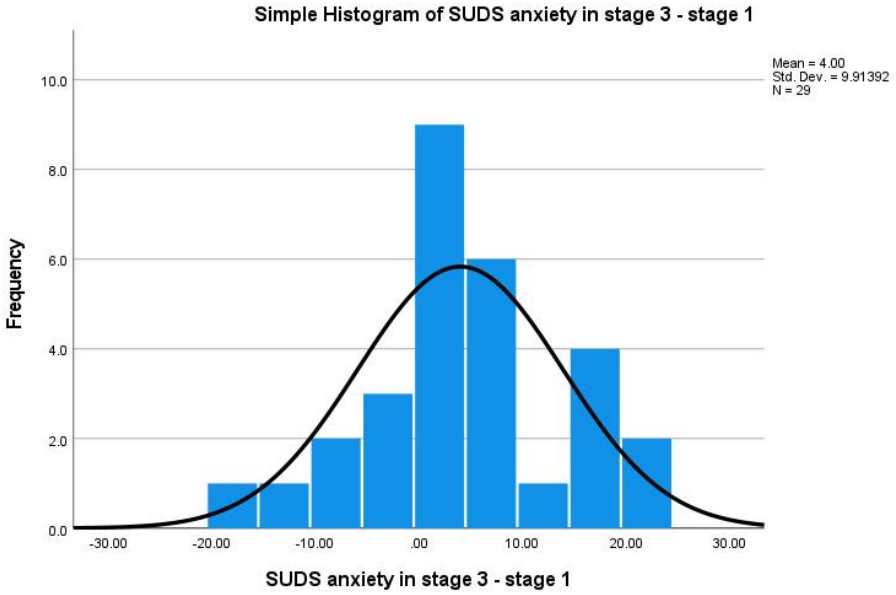


Figure 4.7: Distribution of SUDS increase between stages 1-3

enough to be suitable for Pearson's correlation.

On average, participant SCL increased by 31%. The minimum increase was -49%, and the maximum increase was 100%. Based on the boxplot (Figure 4.9), two participants were outliers, each on the opposite ends of the range. Upon examining the data, these participants had low baseline SCL (.267 for the low end and .932 for the high end). This might indicate that it might be more difficult to observe % changes in participants with low baseline SCL. However, further experiments would be required in order to test this. The outlier values were kept in the dataset. A histogram of SCL increase was drawn to examine the data distribution (Figure 4.10). It can be deduced that the distribution of the data is near normal.

As described on Table 4.4; reaction times ranged between 790.61 and 1257.42. The large range is expected due to the participants' age and VR exposure variance. The mean reaction time was 975.7, and the standard deviation was 124.92. There were no outliers in the data. ESE values ranged between -267.32

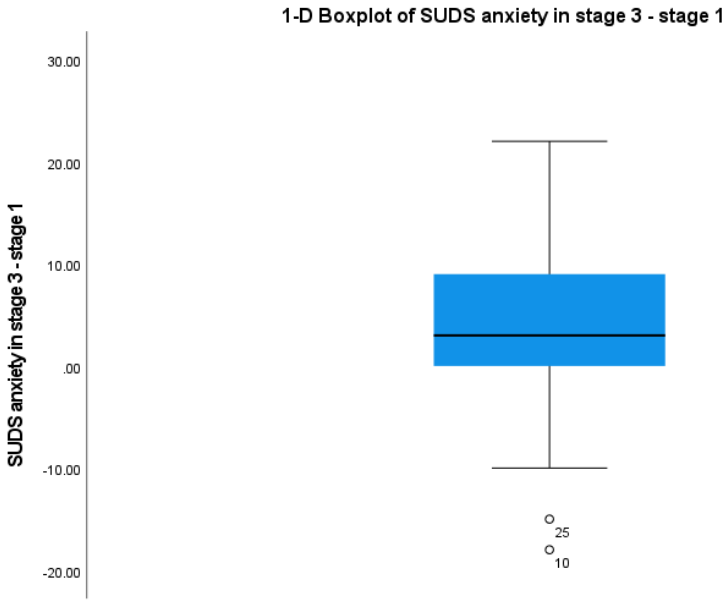


Figure 4.8: Boxplot of SUDS increase

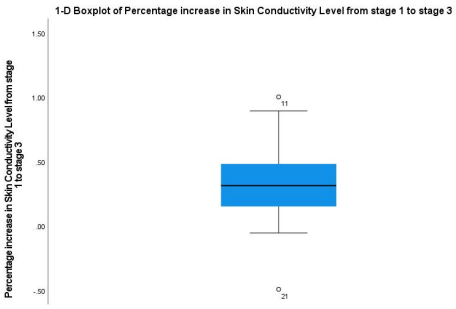


Figure 4.9: Boxplot of % SCL increase

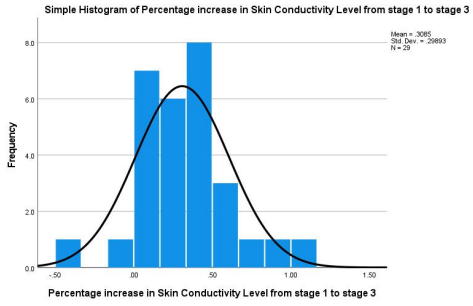


Figure 4.10: Distribution of % SCL increase between stages 1-3

to 172.97. The mean ESE was -29.78, and the standard deviation was 95.46. Upon examining the boxplot, there was one outlier with an ESE of -267.32. This user was also an outlier in terms of SCL increase; their SCL decreased 49% between stages 1 and 3. This outlier was kept in the data. Histogram of the overall reaction time and the reaction time difference between the stages 3-1 can be observed in figures 4.12 and 4.13, respectively. It can be observed from the histograms that these values follow a normal distribution.

4.4 Results

Repeated measures ANOVA was used to compare means between three different stages. ANOVA is used to make meaningful comparisons to spot the mean

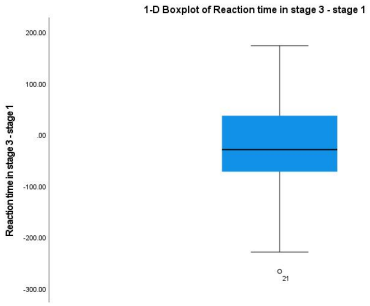


Figure 4.11: Boxplot of % SCL increase

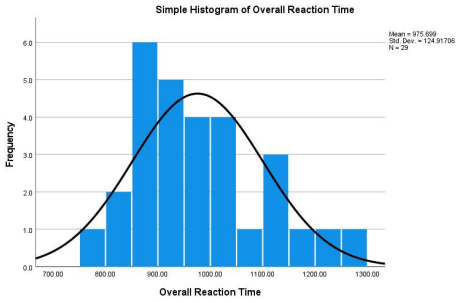


Figure 4.12: Distribution of overall reaction time

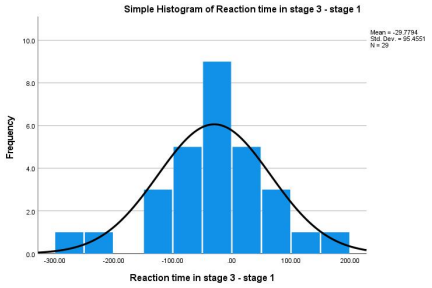


Figure 4.13: Distribution of emotional Stroop effect

difference between more than one group. It stands for analysis of mean variance and is computed by dividing mean squares. The F value is the result of this division.

Repeated measures ANOVA is used to compare means when there are no personal differences between the groups, meaning the same group of participants tested every treatment and condition. It is reliable to measure the impact of the condition/treatment with this test because the group variability and the distributions can be ignored. The sample is matched, so each condition (stages 1,2 and 3) is compared per participant. As the groups are not to be compared and the data is paired, the shape and the distribution of the group are irrelevant.

Repeated Measures ANOVA compares three levels in stages: stages 1, 2 and 3. The means of user SCL (measured in μS), self-reported anxiety (measured by SUDS) and reaction time (measured in ms) were compared between these three levels. The variables used for the comparison were GSR1, GSR2 and GSR3 for SCL; SelfVR1, SelfVR2 and SelfVR3 for self-report difference; and RT1, RT2 and RT3 were reaction time.

Based on Mauchly's test of sphericity, the data was not spherical (Table I.1).

Data sphericity refers to the situation where variances are equal between all combinations. The ANOVA results can be inflated and unreliable if the data is not spherical. As the data violates the sphericity assumption, the probability values must be corrected. Greenhouse-Geisser correction was applied to correct the probability value. Greenhouse-Geisser correction increases the p-value to account for the inflated F value.

Where significant results were found, Bonferroni post hoc corrections were applied to further investigate the relationship between levels. Bonferroni post hoc corrects the probability value by compensating for the increase in p when many comparisons are run simultaneously. As such, the application of this method causes a loss of power. This can be an issue as the number of comparisons increases. Due to the decrease of power, it increases the possibility of a type II error. A type II error refers to accepting a null hypothesis when it is actually false.

4.4.1 Skin Conductivity

The mean levels of SCL between different stages can be viewed in Figure 4.14. On average, a user's anxiety increased 31% (SD = .30) from stages 1 to 3. Based on the repeated measures ANOVA results, there is a significant difference in SCL between the three levels ($F = 20.182$, $p < .001$) (Table I.2). Bonferroni post hoc results show a significant difference between each level (Table I.3).

4.4.2 Self-reported Anxiety

The anxiety scores were measured using SUDS. The difference in means of SUDS between stage 3 and 1 was 4. There was a significant difference in self-reported anxiety between the three levels ($F = 3.877$, $p < .05$) (Table I.2). Bonferroni post hoc results did not reveal any significant pair-wise differences between the stages.

4.4.3 Reaction Time

The difference between each stage in reaction time was investigated using the instructions from Ben-Haim et al. published in 2016 [BHWH+16]. The average reaction time in stage 1 is 980.44ms, the average reaction time in stage 2 is 973.85ms, and the average reaction time in stage 3 is 951.96ms. The users' performance was enhanced in the emotional stage, with an increase of 29.8

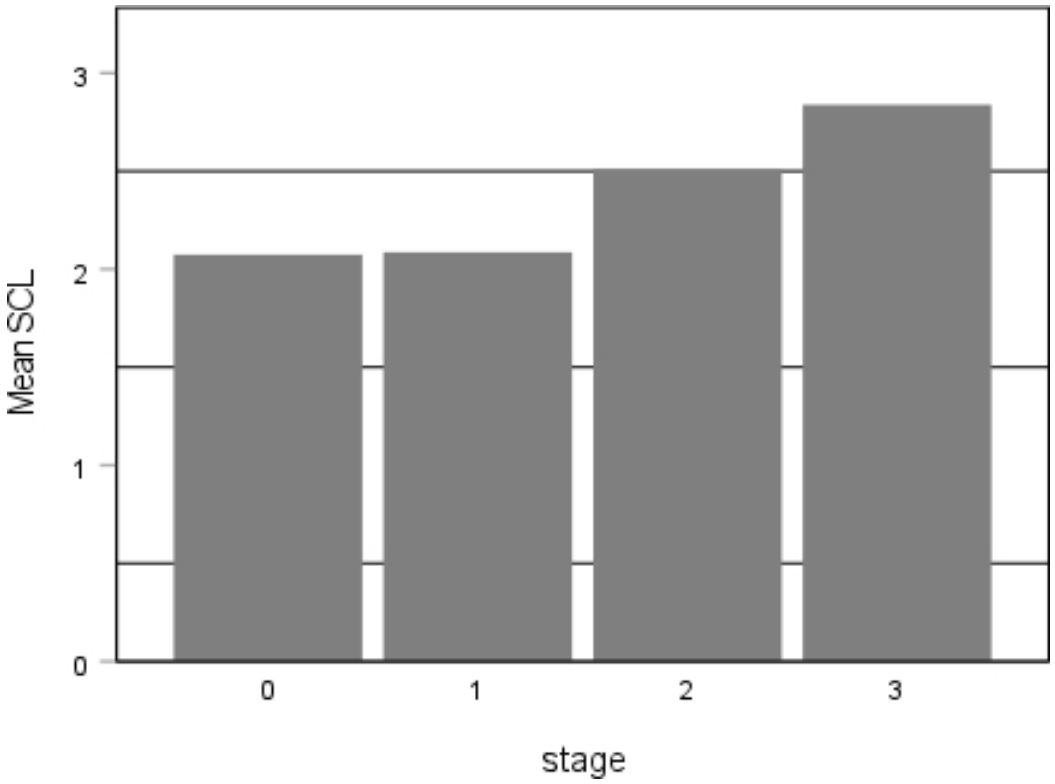


Figure 4.14: Mean SCL for baseline (0), stages 1, 2 and 3

milliseconds in performance.

Repeated measures ANOVA results show no significant difference in reaction time between the three levels ($F = 1.996$, $p > .05$) (Table I.2). However, as discussed in the descriptive statistics section, it should be noted that significant differences in reaction time can be challenging to spot due to lack of power. The vital thing to note is the decrease in reaction time, which is the opposite of the literature. This effect can be viewed across all participants in Figure 4.15.

4.4.4 Other Findings

4.4.4.1 Effect of initial anxiety measured by STAI on performance and user state

As mentioned in the descriptives section, the STAI distribution is not normal. This means that care needs to be taken when correlations are examined. The distribution is heavily skewed to the right (Figure 4.4), with many participants having an STAI score of 0. So, it might be difficult to interpret the effects of initial STAI ratings.

As the STAI distribution was not normal, the Spearman correlation was used

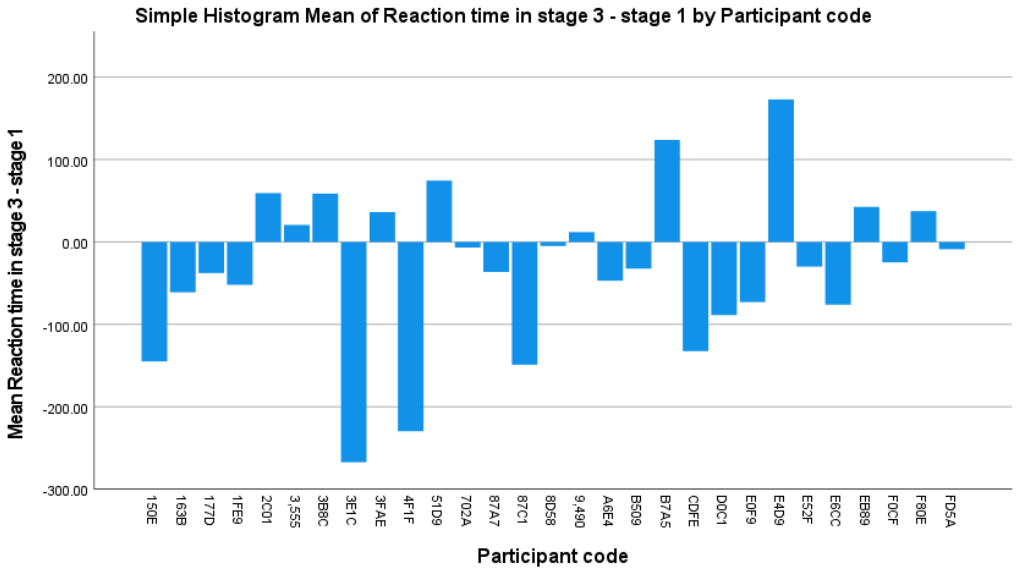


Figure 4.15: ESE per each participant

instead of Pearson. There were no significant relationships found between STAI and either SCL increase, reaction time increase or self-report anxiety increase.

Using Pearson correlation, STAI had a weak linear positive correlation with the difference in SCL between the first and third stages ($r = .457$, $p < .05$), meaning the more anxious the user was initially, the more their SCL increased between the first and third stages. Although this correlation is likely biased due to the heavy skew, it is interesting to look at its effect on type I errors, accepting alternative hypotheses when they are wrong. The relationship can also be observed in the scatterplot (Figure 4.16).

4.4.4.2 Relationship between different anxiety measures and reaction time

Upon drawing and examining scatterplots in SPSS, no relationships to be examined can be observed. Upon running formal Pearson's correlation tests, it can be confirmed that no linear relationship exists.

The scatterplot between the percent increase in SCL and ESE can be viewed in Figure 4.17. It seems that the higher the skin conductivity of the user, the more their reaction time decreases. This might be related to the Yerkes-Dodson effect, as well as a learning effect. Upon running Pearson's correlation, no linear relationship was found ($r = -0.47$, $p > .05$).

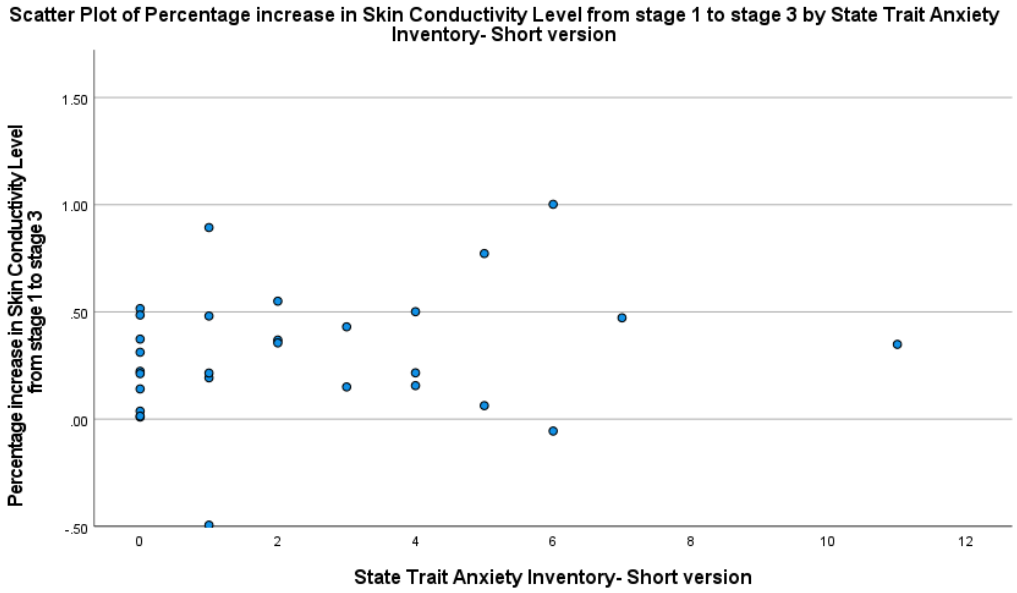


Figure 4.16: Scatterplot between initial STAI scores and SCL increase

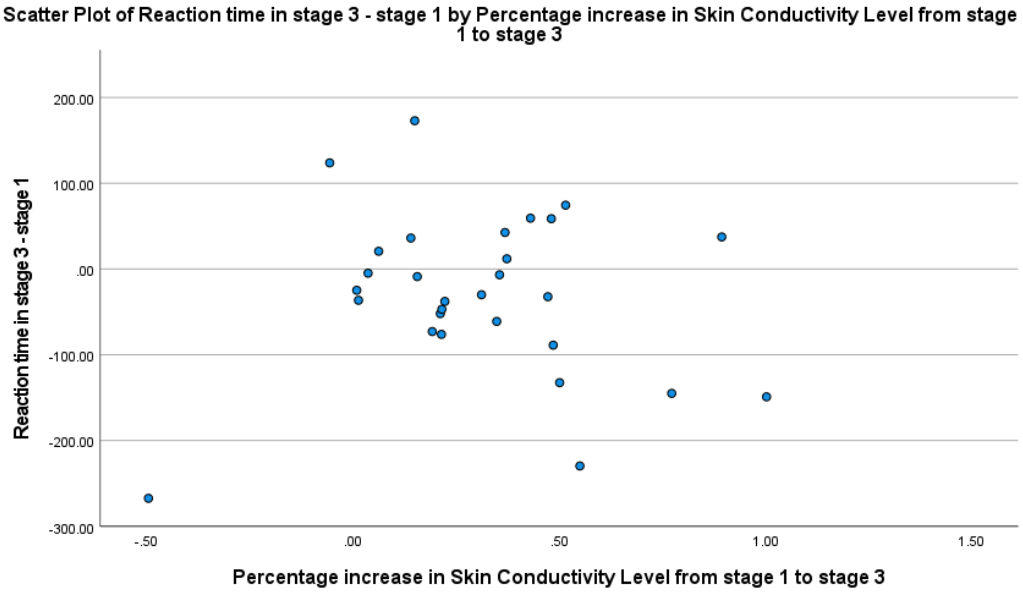


Figure 4.17: Scatterplot between SCL percentage increase and ESE

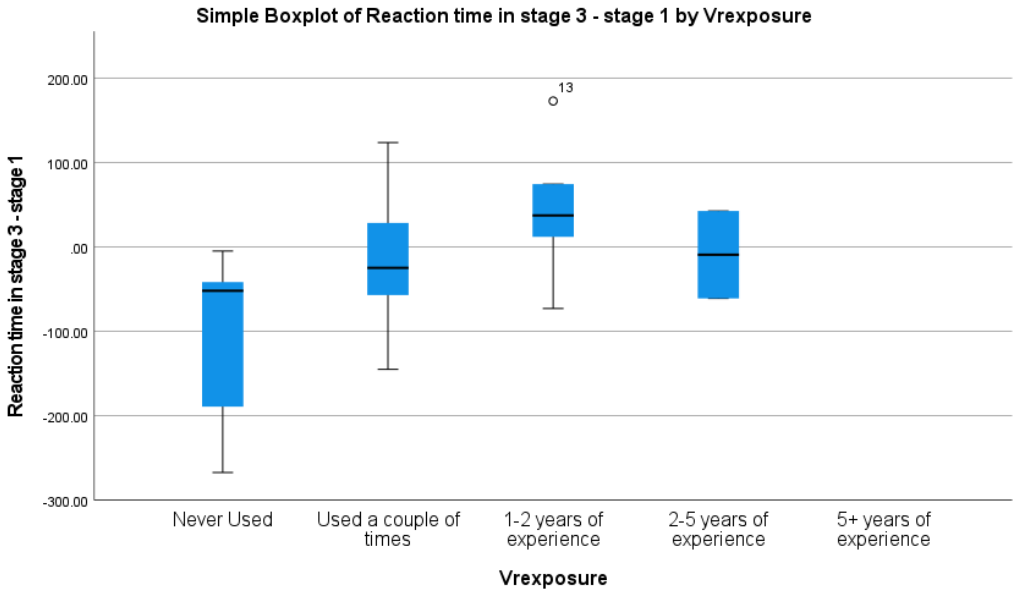


Figure 4.18: Boxplot of mean ESE by VR exposure

4.4.4.3 ANOVA for demographic effects

The information in the descriptive statistics section was used to further examine means where there might be a significant effect. ANOVA tests were run to find the effects of age group and VR exposure on overall reaction time. The F value of the age group on reaction time was 2.80, making the probability 0.6, which is not significant. No other effects of age group were found. It must be noted that most of the participants belonged to two age groups (18-24 and 25-35), and there were two or fewer participants per other group, which meant it was not possible to apply Bonferroni correction to adjust power.

There was a significant effect of VR exposure on ESE ($F = 3.67$, $p > .5$). Furthermore, Bonferroni post hoc results reveal that there was a significant pairwise difference between participants who never used VR and participants with 1-2 years of experience in VR. The mean difference between participants with 1-2 years of experience and participants who never used VR was 157.08 ($p < .05$). A boxplot of mean ESE by VR exposure can be viewed in Figure 4.18 This might signify that the emotional Stroop effect was more pronounced for experienced VR users. It should be kept in mind that there were only a few participants per group due to sample size, which might make ANOVA comparisons a bit more difficult by decreasing the power.

4.4.4.4 Comments

Open-ended comments from the participants were grouped into some common categories. In terms of the experience, two participants commented that the platform was too low for them to reach, two participants commented that it was easy to misclick due to the sensitivity of the sensors and the interactions system, and one participant commented that more feedback was necessary in terms of performance and progress, one participant commented that it was challenging to quantify one's own anxiety and one user commented that they had a lot of fun during the experiment. In terms of the equipment used, two participants commented that the HMD was very heavy, two participants commented that the HMD was not tight enough around their head, two participants commented on not being able to use their prescription glasses with the HMD, two participants commented that the earphones were too loose, one participant commented that the volume was too high and one participant commented that the equipment needs to be reduced for comfort.

4.5 Discussion

This chapter analysed the new application, eVRST, and compared it to previous in vivo applications. As opposed to previous versions of eStroop, this application includes three conditions instead of two. Therefore, it can be challenging to make direct comparisons between the previous in vivo and virtual reality versions, including two conditions. Furthermore, this thesis focussed on the change of arousal throughout different conditions instead of solely focussing on the difference in performance. Traditional Stroop task is created to serve as a measure of cognitive ability. Therefore, previous experiments mainly focus on performance and the so-called "Stroop effect" [KBB96, ST17]. Conversely, in this experiment, it was aimed to record the trends of arousal change between the conditions rather than performance differences.

Although it was not the main objective of the study, ESE was examined. Interestingly, the reaction time difference between the neutral (condition 1) and high emotional conditions (condition 3) was in contrast with previous literature [GWM+19, ST17]; the reaction time decreased between the first and third conditions. A block design was used, with emotional conditions following neutral conditions, to enhance the effect of ESE as suggested by Ben-Haim et al [BHWH+16]. Based on prior research, using a random order for emotional

Stroop diminishes the emotional effect because the emotions from the previous conditions impact the current mental state of the participants.

Despite the possibility of fatigue, the performance of the participants increased in the last condition. This result is in contrast with previous work on the Stroop effect. It is possible that the increase in performance was caused by the learning effect, increased by the difficulty of getting used to a task in Virtual Reality as compared to in vivo and computer applications. This idea can be supported by the finding that VR experience affects ESE, with more experienced users experiencing higher ESE. In future applications for finding ESE, it is important to prolong tutorials to eliminate learning effects while keeping breaks significant to minimise fatigue.

In the experimental results, there is a clear trend of increased arousal through stages, as evidenced by the users' SCLs. While we have strong evidence pointing to increased arousal, the valence is interpreted as being negative based on the responses to self-reported anxiety. However, the degree to which the participants were affected emotionally is difficult to interpret because, while significant, the difference between conditions in self-reported anxiety was not very high. The small difference in self-reported anxiety might suggest the possibility of impacts of different factors on the increased arousal; such as excitement of the task. Additionally, the ESE was not significant.

The results show that the eVRST application is effective at showing different levels of anxiety; however, based on self-reported responses, the difference is mild. Whether the classification is possible using this dataset is difficult as the changes in anxiety might be too subtle to be used as part of a more stressful environment. Machine learning will be utilised to inquire whether it is possible to accurately differentiate between the conditions of this experiment. Since these changes are subtle, machine learning models being able to differentiate between these levels of anxiety would mean they would perform even better when discriminating between more pronounced changes. While applying machine-learning techniques, the results of the extensive scoping review in Chapter 2 will be used for background and architecture.

Based on the responses to the usability questionnaire, most users found the application easy to understand the instructions and carry on the given task using the tutorial. Despite the efforts to decrease the amount of equipment used, it still had a slight effect on the immersion in the environment and comfort. Based on some participant feedback, this could partly be attributed to

the weight of the HMD used. The experience caused minimal motion sickness. Most users were highly engaged during the experiment and not very bored. A possible change to increase usability could be to use a different HMD, as it was deemed too heavy by several users.

4.5.1 Limitations

This study has some limitations based on the sample and the difficulty of dealing with subjective human emotions. First of all, the study was conducted in a large university, which possibly biased the sample to be more educated and technically able. 76% of the participants had tried VR before, which is higher than the population average [Ins20]. It is possible that the experience of the sample compared to the general population might have biased the results and prevented the discovery of unique challenges that might be caused by being new to VR.

Another limitation is that in order to amplify the participants' arousal in the stage, the different groups were presented in the same order to every participant. This makes it easier to gather data in different states of emotion; however, it makes it harder to derive information about the specific reason for the increased arousal and Stroop effect. Fatigue and prolonged VR exposure or VR sickness might have contributed to these emotions.

4.5.2 Recommendations

This experiment outlines the importance of choosing the correct study design based on the goal. The order of blocks of different conditions in eStroop significantly affects the anxiety measured in that participant. If the effects of the words themselves need to be studied, the order of conditions must be permuted and randomised. However, sufficient time must be given between conditions so that the effects of the previous condition subsides. Ideally, the conditions would be administered in separate sessions. If the aim is to use the environment as a stressor, it is important to order them from mild to severe, as in this experiment.

The results show that the effect of ESE is negligable. There are multiple possible reasons for a nonsignificant ESE. Since there is no cognitive interference involved in the emotional version of the task, the performance is not as negatively impacted by the changes in conditions. eStroop does not

increase the cognitive load at the same intensity as the traditional Stroop task. This means that its effect on arousal is not as intense. The decreased Stroop effect between emotional and traditional versions of the task has been well-documented [ACL04]. However, since this experiment aims to measure emotional anxiety, then the effect being emotional and non-cognitive is intentional; because it means that the results are not contaminated by the task difficulty and cognitive load.

The cognitive load in the eStroop can be increased by introducing time pressure or distractors. If the goal is to strengthen the time delay effect of Stroop, these adjustments can bring up the difficulty and complexity.

A more extensive tutorial is required for future Virtual Reality applications of eStroop. Depending on the population and their technical abilities, it might be essential to teach the controls of VR and detail the task.

4.6 Conclusion

This chapter details verifying and validating a novel platform for administering eStroop in Virtual Reality. The eVRST application is effective in delivering differing levels of anxiety per condition, as evidenced by the self-report and physiological analysis. Consequently, the data collected from this experiment can be used for further machine learning analysis. The following chapter will further analyse the importance of heart rate, skin conductance and electrical brain activity in distinguishing between no anxiety, mild anxiety and severe anxiety conditions.

Chapter 5

Anxiety Classification

This chapter uses state of the art machine learning classification methods to analyse anxiety between conditions in the eVRST environment and validate such methods. As outlined in the introduction, the main goal of this thesis is the real-time classification of anxiety. After its validation in the eVRST experiment, emotional Stroop conditions are used as a means to label the data for prediction. Based on the collected information, it is essential to accurately classify the condition (neutral, mild, severe) in the specified timeframe.

Explainability of the algorithm is prioritised as the application is health-related; it is of utmost importance to understand what the algorithm uses to detect anxiety. Thus, black-box approaches like ANNs will only be used if substantial performance differences exist. As the classification will be administered in real-time, efficiency and performance are also of high priority. The challenge of this validation lies in dealing with the noise in wearable technology data and personal differences. This chapter extends the publication entitled "Real-time Classification of Anxiety in Virtual Reality Therapy Using Biosensors and a Convolutional Neural Network", published in the MDPI Biosensors journal in 2024 [MTM24].

5.1 Background

Machine learning is a very trendy topic, with an unsurmountable amount of new research constantly being conducted. Based on records on the Scopus database, records with the keyword "machine learning" more than doubled in the last five years. In 2018, 0.77% of all records had machine learning as a

keyword. In contrast, 2023 had 2.54% of all documents. Figure 5.1 shows the trends of records which include the "machine learning" keyword in Scopus.

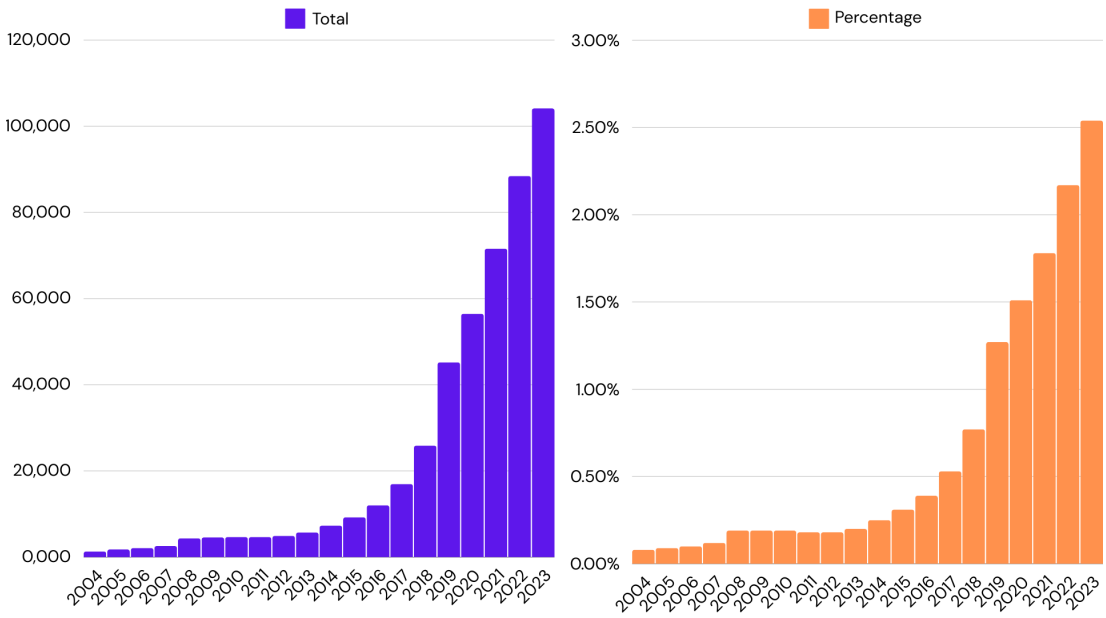


Figure 5.1: Left: total number of records with the keyword machine learning. Right: The proportion of machine learning records is compared to all records published in the same year.

This boom in popularity is due to the usefulness of machine learning in various fields, ranging from medicine to agriculture and business. In medicine, for instance, machine learning algorithms are employed to predict disease outbreaks, diagnose illnesses from medical imaging, and personalise treatment plans based on patient history. In agriculture, machine learning can aid in optimising crop yields through predictive analytics and automated systems for monitoring soil and plant health. In business, machine learning is utilised for customer analysis, fraud detection, and improving supply chain efficiency.

These rapid advancements in machine learning methods and techniques are a response to the evolving needs of industry and academia. One of the critical areas of innovation is the development of new strategies, such as natural language processors and transformers. These advancements have improved tasks such as language translation, sentiment analysis, and text generation. As the technical space progresses, it is vital to pay attention to qualitative aspects of machine learning, such as ensuring unbiased databases and addressing ethical concerns. Bias in machine learning models can lead to unfair treatment and strengthen existing inequalities, making developing techniques for bias detection and mitigation essential.

Another critical aspect is the explainability of machine learning models. As models become more complex, especially with deep learning and neural networks, understanding how models make decisions becomes increasingly difficult. The black-box nature of advanced models poses challenges for their adoption in fields where accountability and transparency are crucial, such as healthcare and finance.

There are two primary types of machine learning: supervised and unsupervised. Supervised machine learning refers to algorithms trained using pre-labelled data, including input data and the corresponding output. This type of learning is commonly used in applications where the goal is to predict a specific outcome based on historical data. For instance, in supervised learning, the algorithm learns from a dataset containing input-output pairs, such as images of handwritten digits and their labels. Once trained, the algorithm can accurately predict the label of unseen images. This approach is beneficial in applications like image recognition, speech recognition, and medical diagnosis, where labelled data is available, and the objective is to classify new examples correctly. The most common type of learning used in emotion detection is supervised learning. In emotion detection, supervised learning algorithms are trained on labelled datasets where the input data consists of features extracted from text, audio, or video, and the output labels represent different emotions, such as happiness, sadness, anger, or fear.

In addition to classification, supervised learning also includes regression tasks. Regression is used when the output variable is a continuous value rather than a label. An example of regression would be predicting the market value of a house based on features such as house size, location, and the number of bedrooms. The algorithm learns the relationship between the input features and the continuous output variable from the training data and can then predict the value of a new house given its features. Regression models are widely used in financial forecasting, such as predicting stock prices and assessing the impact of marketing campaigns on sales. The advantage of supervised learning is that it provides a clear framework for training models to make accurate predictions based on historical data.

Unsupervised machine learning, on the other hand, is used to gain insights from unlabeled data where there is no predetermined output. In unsupervised learning, the algorithm explores the data's underlying structure without prior knowledge of the outcomes. A common type of unsupervised learning is

clustering, where data points are grouped based on their similarities. Clustering is widely used in various domains, including market research, social network analysis, and image segmentation. In market research, clustering can help identify distinct customer segments based on purchasing behaviour, enabling businesses to tailor their marketing strategies to different groups. In social network analysis, clustering can reveal communities within a network, such as groups of users who frequently interact with each other. In image segmentation, clustering algorithms can partition an image into regions with similar characteristics, aiding tasks such as object detection and image recognition.

5.1.1 Related Work

Significant advancements in the field of anxiety classification have occurred in recent years. While many studies focus on using wearables for continuous anxiety prediction, the applications of such studies in VR are limited. Applying anxiety prediction in VR can be challenging due to ergonomic challenges. Mobility is vital in VR, which can introduce noise into signals. Similarly, the HMD can negatively affect the electrical signal quality of some head-based wearables such as EEG.

Classification is the most commonly used method for predicting anxiety using machine learning. In this type of algorithm, deciding how many outputs are required is necessary. Researchers often seek a balance where precision is improved without sacrificing the algorithm's accuracy. The most common levels of anxiety measured with this method range from two outputs (e.g., no anxiety and anxiety) to four outputs (e.g., no anxiety, low anxiety, mild anxiety, and severe anxiety). Data is labelled using predefined conditions such as established tasks or content and self-report measurements like the STAI-Y1.

Some researchers have used a regression approach with self-report measures as the ground truth [PPM+20]. This results in a continuous score, which is then placed in predefined classes based on the continuous score. While black-box approaches like neural networks are generally avoided, some research has utilised them with high accuracy results [BMM+20]. Despite their complexity, Neural networks can capture intricate patterns in data, making them suitable for tasks where traditional models fall short.

Integrating multimodal data, such as combining physiological signals with behavioural data, is also gaining traction. This approach can provide a more comprehensive understanding of anxiety levels by capturing different aspects of

the anxiety response. For example, combining heart rate variability with facial expression analysis can improve the accuracy and robustness of anxiety detection systems. Researchers are exploring various sensor technologies and data fusion techniques to enhance the reliability of these models.

A study by Šalkevicius et al. [vDML19] used signals such as EDA, BVP, and SKT to classify anxiety levels within VRET. Their work presents their cloud-based framework for tailoring anxiety classification in VR. Physiological data was collected from 30 participants, and the resulting dataset was used to train a machine-learning model. Researchers used public speaking anxiety in VR as the basis of their exposure and stress stimulus, and they employed four levels of output. These output levels were low, mild, moderate, and high anxiety. They achieved an 80% cross-participant accuracy with an SVM, validated using the 10-fold cross-validation method.

Similarly, Petrescu et al. [PPM+20] used HR and EDA features with a regression model and a rule-based system to extract classes from a regression model. They used height as the stressing stimulus. They achieved 92% accuracy among seven participants. However, the authors note that the overall accuracy might be biased due to the uneven distribution of classes in their data. They report that the accuracy for the mild anxiety class was 81% and the high anxiety class was 76%, based on a confusion matrix. It should also be noted that the effect of interpersonal differences is diminished with a small sample size of seven participants. This research highlights the importance of addressing class imbalance in machine learning models to ensure fair and accurate predictions.

Cho et al. [CHO+17] took a slightly different approach and combined arithmetic tests with VR environment stimuli to elicit the stress response. The authors used HRV, SKT and EDA to estimate anxiety levels. Twenty individuals have participated in the study. Through the augmentation of their "integrated feature" (IT = HRV + SC + SKT), their K-ELM algorithm was able to differentiate between their two baseline environments (before and after the stimulus) and three anxiety environments (mild, moderate, and severe) with 95% accuracy. This study demonstrates the efficacy of feature engineering for improving the performance of machine learning models.

Current work focuses on both the comfort and accuracy of the system to seamlessly and objectively measure the anxiety level of the patient, distinguishing between no anxiety, mild anxiety, and severe anxiety. The aim is to create more reliable and tailored anxiety detection and classification methods,

contributing to improved mental health interventions. Integrating machine learning in mental health assessments promises to enhance the precision and personalisation of treatments, ultimately improving patient outcomes.

5.2 Methods

5.2.1 Dataset Creation

As detailed in the previous chapter, the information collection in this experiment was done using the official APIs of the Shimmer and NeuroSky devices. As such, it already includes some real-time processing, which is very useful for the real-time application. In the case of the PPG, the values received already go through R-R peak detection, outputting the resulting heart-rate value. More information on this can be found in the Shimmer validation paper [BDG+10].

Using the Shimmer API, a low-pass filter with a corner frequency of 5 Hz was applied to the PPG signal. Similarly, a high-pass filter was used to filter frequencies above 0.5 Hz. The PPG to heart rate calculations were done using the filtered signals. These can be viewed in Shimmer documentation [Eng22].

In GSR signals, the noise is generally contained in high frequencies. Therefore, only a real-time low-pass filter was applied with a corner frequency of 5 Hz, as per Shimmer recommendations [Eng22]. The low-pass filter used was Finite Impulse Response (FIR) Blackman-windowed-sinc.

All the preparation for the EEG signals was done using the NeuroSky software. The corresponding frequencies for each band are as follows: delta; 1-3 Hz, theta; 4-7 Hz, alpha1; 8-9 Hz, alpha2; 10-12 Hz, beta1; 13-17 Hz, beta2; 18-3 Hz, gamma1; 31-40 Hz and gamma2; 41-50 Hz. An ”e-sense” technology was used to compare each band every second. More information can be found in the Neurosky documentation [Neu22b].

The values from before the start of the experiment and after the end of the experiment were manually removed using timestamps. All the separate .csv files for experiments were processed and combined into a single .csv file.

5.2.2 Data Preparation

5.2.2.1 Validation Splitting

A simple method of train-test split was applied to compare different classification methods. The dataset was split into 80% training data and 20% testing data using the GroupShuffleSplit method from the SKlearn library. This method randomly splits data based on groups provided, in this case, the participant ID. The return values are the indices of the selected IDs on the data frame in the form of arrays. These indices are then used to declare train and validation data frames.

5.2.2.2 Normalisation

Normalisation was applied to ANN, SVM, KNN, and CNN. The method used for normalisation is (value - mean) / std. It should be noted that the baseline-adjusted values are redundant in the case of normalised datasets, such as in the case of Neural Networks. Therefore, the baseline values were removed for these models, leaving only 15 features instead of 23.

5.2.2.3 Windowing

Windowing was applied separately for the CNN. All models except CNN used rolling means and standard deviations with window sizes of 3 s, 5 s, 10 s, and 30 s and step sizes of 1 s, 1 s, 2 s and 10 s, respectively. Methods from NumPy library, numpy.roll(), numpy.mean() and numpy.std() were used for these calculations. Baseline adjusted values were calculated by dividing each value by the mean of baseline recordings of the same metric.

For the CNN, the dataset was iterated in window sizes of 3 s, 5 s, 10 s, and 30 s and step sizes of 1 s, 1 s, 2 s and 10 s, respectively. The raw values of window size $\times$ features were used as input. Each matrix of window size $\times$ features shape was labelled with the correct condition.

5.2.3 Feature Extraction

First of all, relevant features were identified from previous literature. As the data is time-series data, all metrics' mean values and standard deviation over the sliding window were calculated. Since every person has different reactions and baseline values, mean values adjusted by baseline were also included in the

features. These were applied to the HR, SCL, and EEG band wave values. The resulting features can be viewed in Table 5.1.

Table 5.1: The extracted features

SCL	SCL_mean
	SCL_baseline
	SCL_SD
HR	HR_mean
	HR_baseline
	HR_SD
EEG	delta_mean
	delta_baseline
	delta_SD
	theta_mean
	theta_baseline
	theta_SD
	lowalpha_mean
	lowalpha_baseline
	lowalpha_SD
	highalpha_mean
	highalpha_baseline
	highalpha_SD
	lowbeta_mean
	lowbeta_baseline
	lowbeta_SD
	highbeta_mean
	highbeta_baseline
	highbeta_SD
	lowgamma_mean
	lowgamma_baseline
	lowgamma_SD
	highgamma_mean
	highgamma_baseline
	highgamma_SD

Adding minimum and maximum values was considered but decided against due to the short nature of the time windows and the results of direct experimentation with their inclusion.

As the input for the CNN is different, different feature extraction was applied. Instead of getting mean, baseline adjusted mean, and standard variation over the course of the time window for the metrics, all of the metrics containing the row were used, forming matrices in the shape of the time window $\times$ metrics. In this case, the extracted features were GSR, HR, delta, theta, low alpha, high

alpha, low beta, high beta, low gamma, and high gamma.

The GSR data was deconvolved into phasic and tonic components by applying CDA using the NeuroKit library for Python [MPL+21]. It is hypothesised that the GSR signal will be comparable to the deconvolved signals, so the additional signals are also analysed in the end.

5.2.4 Feature Selection

5.2.4.1 ANOVA

An ANOVA test was used to determine the relationship between each feature and the conditions to identify the most valuable features. The features with non-significant F values were discarded. These features were HR standard deviation, delta mean, delta mean adjusted by baseline, delta standard deviation, high alpha mean adjusted by baseline, high alpha standard deviation, and low gamma standard deviation. Features and their F values can be viewed in Table 5.2.

In the case of the CNN, the selected features were GSR, HR, theta, low alpha, low beta, high beta, low gamma, and high gamma. Delta and high alpha were eliminated due to no relationship with the condition. The F values for the CNN features can be seen in Table 5.3.

5.2.4.2 Gini Index

To have a better view of the feature importance, a Random Decision Forest was constructed and the Gini index was used to plot the importance of each feature. The Gini index measures the mean decrease in impurity and shows the impact each feature has on the final decision. The feature importances of the RDF, and feature importances in each tree were calculated. The aggregated results, including the standard deviation per tree, can be seen in Figure 5.2.

As can be seen from the Gini importance, the ranking of feature importances is similar to those arising from the ANOVA results. GSR features have the highest importance, with GSR in relation to baseline being the most important, mean being the second, and standard deviation being the third. They are followed by mean heart rate and heart rate in relation to baseline. EEG-related features lack importance, with the most important ones being gamma1 in relation to baseline. This is in contrast to ANOVA results, where beta2 in relation to baseline was of the highest importance. Delta also has high importance

Table 5.2: The relationship between the feature and the condition

Feature	**F**
GSR_baseline	2379.1***
GSR_mean	115.3***
HR_base	18.4***
beta2_base	17.3***
beta2_mean	15.4***
HR_mean	15.2***
alpha1_mean	10.9***
gamma1_mean	9.8***
alpha1_base	9.0***
gamma2_base	8.8***
gamma2_mean	7.8***
beta2_SD	7.2***
theta_mean	7.1***
beta1_mean	6.2***
alpha2_mean	5.8**
beta1_base	4.7**
gamma2_SD	4.6**
alpha1_SD	4.5*
gamma1_base	4.2*
theta_base	4.0*
gsr_SD	3.9*
theta_SD	3.6*
beta1_SD	3.1*
alpha2_base	2.8
delta_mean	2.4
delta_base	1.8
gamma1_SD	1.7
delta_SD	1.2
alpha2_SD	1.1
hr_SD	0.5

*** $p < .001$, ** $p < .01$, * $p < .05$

compared to the results of the ANOVA, which might suggest a more complex interaction between delta and anxiety that cannot be measured using ANOVA.

5.2.5 Feature Convergence

Correlations between each of the metrics were computed to examine the relationship between each metric. The correlation matrix can be viewed in Figure 5.3. Note that the values are normalised before correlations.

As can be seen from the correlation matrix, there is a strong positive internal

Table 5.3: The relationship between the CNN features and the condition

Feature	F
GSR	117.22***
HR	14.75***
beta2	7.42***
gamma1	5.03***
alpha1	4.67***
gamma2	3.96**
theta	3.13***
beta1	3.06**
alpha2	2.74
delta	0.81

*** p < .001, ** p < .01, * p < .05

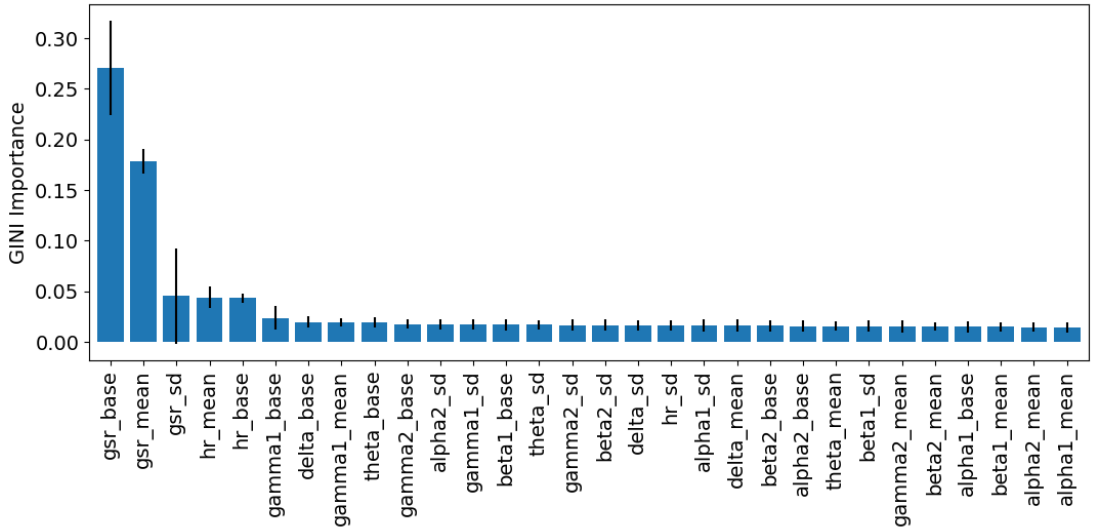


Figure 5.2: Gini importances

correlation among EEG bands. Excluding delta, the lowest correlation is 0.4 between gamma2 and theta, and the highest is 0.8. between gamma2 and beta2. Delta correlations with other EEG frequencies range between 0.16 (gamma2) and 0.42 (theta). The SCL is highly correlated with GSR, as hypothesised. SCR is not strongly correlated with any values. There are no strong correlations between any of the separate metrics. The correlation between heart rate and GSR is 0.12.

	alpha1	alpha2	beta1	beta2	delta	theta	gamma1	gamma2	GSR	heartrate	EDA_Tonic	EDA_Phasic
alpha1	1.00	0.62	0.59	0.62	0.34	0.54	0.59	0.52	0.00	-0.06	0.00	0.01
alpha2	0.62	1.00	0.65	0.67	0.34	0.56	0.65	0.62	-0.00	-0.07	-0.00	0.01
beta1	0.59	0.65	1.00	0.79	0.30	0.50	0.68	0.73	0.00	-0.10	0.00	0.00
beta2	0.62	0.67	0.79	1.00	0.31	0.55	0.79	0.80	-0.02	-0.09	-0.02	0.01
delta	0.34	0.34	0.30	0.31	1.00	0.42	0.32	0.18	0.05	0.00	0.05	-0.00
theta	0.54	0.56	0.50	0.55	0.42	1.00	0.58	0.40	-0.01	-0.05	-0.01	0.02
gamma1	0.59	0.65	0.68	0.79	0.32	0.58	1.00	0.74	-0.02	-0.10	-0.02	0.01
gamma2	0.52	0.62	0.73	0.80	0.18	0.40	0.74	1.00	-0.03	-0.11	-0.03	0.01
GSR	0.00	-0.00	0.00	-0.02	0.05	-0.01	-0.02	-0.03	1.00	0.12	1.00	0.05
heartrate	-0.06	-0.07	-0.10	-0.09	0.00	-0.05	-0.10	-0.11	0.12	1.00	0.12	0.01
EDA_Tonic	0.00	-0.00	0.00	-0.02	0.05	-0.01	-0.02	-0.03	1.00	0.12	1.00	0.02
EDA_Phasic	0.01	0.01	0.00	0.01	-0.00	0.02	0.01	0.01	0.05	0.01	0.02	1.00

Figure 5.3: Correlation Matrix between each feature

5.3 Results

Multiple models and window sizes were tested based on the properties data and prior literature.

5.3.1 Models

For model comparison, 3 s time windows were used. Test-train split of 20-80 was applied, and the same test and train groups were used to test each model. A comparison of each model can be seen in Table 5.4.

Table 5.4: Baseline accuracies for each classifier

Classifier	Baseline Accuracy
RDF	52.87%
DT	50.67%
XGB	49.62%
KNN	61.92%
SVM	63.12%
ANN	72.24%

Note: ANN: Artificial Neural Network, DT: Decision Tree, KNN: K-nearest Neighbor, RDF: Random Decision Forest, SVM: Support Vector Machine, XGB: Extreme Gradient Boosting

5.3.1.1 Decision Trees

The first tree model was a simple Binary Tree. For baseline comparisons, parameters with twenty-four features and a maximum of twenty leaf nodes were

used. The tree method was imported from the SKlearn library. When trained on the training set, it had 51% accuracy on the test set.

An RDF was constructed next. The default hyperparameters of the SKlearn RandomForestClassifier method were used, with a maximum tree count of 100. An accuracy of 53% was achieved on the test set. Finally, an XGBoost classifier was used from the XGBoost library. With default hyperparameters, an accuracy of 50% was achieved.

5.3.1.2 Rule-based system

A rule-based if and else function was created based on a Decision Tree. This function was created by first constructing a decision tree with 51% accuracy. This tree was then converted into if and else statements node by node. Creating the rule-based system helps better understand and explain the algorithm. To achieve this, a function that recurses through the tree node by node and turns the nodes into if and else statements were written. If the statement is true, it carries onto the right tree node; if it is false, it goes to the left node using the else statement. The _tree object from the SKlearn library was used to iterate through binary tree nodes that were trained earlier.

Although other algorithms have reached a higher accuracy, this solution was implemented for ease of understanding and computational simplicity. The rule-based if and else statements can be viewed in Snippet 5.1.

```
def predict(gsr_mean, gsr_base, hr_mean, hr_base):
    if gsr_base <= 1.13:
        if gsr_mean <= 0.16:
            return [[  0.,  0.,116.]]
        else:  # if gsr_mean > 0.16
            if gsr_base <= 1.0:
                if gsr_mean <= 0.27:
                    return [[  0.,108.,  0.]]
                else:  # if gsr_mean > 0.27
                    if gsr_base <= 0.86:
                        return [[858., 23.,  0.]]
                    else:  # if gsr_base > 0.86
                        if gsr_mean <= 1.89:
                            if gsr_mean <= 1.33:
                                return [[138.,122.,  0.]]
```

```
                        else:  # if gsr_mean > 1.33
                            return [[ 0.,98., 0.]]
                    else:  # if gsr_mean > 1.89
                        return [[469., 43.,  0.]]
            else:  # if gsr_base > 1.0
                if gsr_mean <= 0.99:
                    return [[76., 6., 0.]]
                else:  # if gsr_mean > 0.99
                    if gsr_mean <= 1.94:
                        return [[ 20.,217.,  3.]]
                    else:  # if gsr_mean > 1.94
                        if gsr_mean <= 2.27:
                            return [[ 28., 23.,103.]]
                        else:  # if gsr_mean > 2.27
                            if gsr_mean <= 5.0:
                                return [[228.,130., 12.]]
                            else:  # if gsr_mean > 5.0
                                return [[ 4.,82., 0.]]
    else:  # if gsr_base > 1.13
        if gsr_mean <= 0.88:
            return [[173.,  0.,  0.]]
        else:  # if gsr_mean > 0.88
            if gsr_base <= 1.35:
                if gsr_mean <= 1.28:
                    return [[  0.,176.,  3.]]
                else:  # if gsr_mean > 1.28
                    if gsr_mean <= 4.03:
                        if hr_mean <= 87.17:
                            if hr_base <= 0.89:
                                return [[ 0.,66.,40.]]
                            else:  # if hr_base > 0.89
                                return [[  4., 20.,167.]]
                        else:  # if hr_mean > 87.17
                            if hr_mean <= 134.83:
                                return [[ 95.,566.,294.]]
                            else:  # if hr_mean > 134.83
                                return [[  4., 36.,136.]]
```

```
                else:  # if gsr_mean > 4.03
                    return [[  0., 14.,147.]]
            else:  # if gsr_base > 1.35
                if gsr_mean <= 1.13:
                    return [[ 0.,53., 0.]]
                else:  # if gsr_mean > 1.13
                    return [[  0., 50.,814.]]
```

Snippet 5.1: Rule-based system

In the code above, the list output with three integers represents the weight (probability) of each outcome: no anxiety (0), mild anxiety (1) and severe anxiety (2).

5.3.1.3 K-Nearest Neighbour

The KNeighborsClassifier method from the SKLearn library was utilised. The default hyperparameters were maintained, with the number of neighbours set to 3. The training resulted in an accuracy of 62% on the test set.

5.3.1.4 Support Vector Machine

The SVM method from the SKLearn library was employed for creating the Support Vector Machine. The kernel is poly, and regularisation is 100. For the other hyperparameters, the default values were used. The training resulted in an accuracy of 63% on the test set.

5.3.1.5 Artificial Neural Network

Considering that the dataset is small with all numerical values, a shallow Neural Network was created. The Keras Sequential model was employed for training the Neural Network. The input layer has 64 neurons and uses activation reLU. It includes three dense layers with ten neurons each, and the reLU function was used for activation. For the output layer, the softmax function was utilised. The selected hyperparameters were categorical_crossentropy for the loss function, Adam for optimiser and categorical_accuracy for metrics. The dataset was trained for twenty epochs with a batch size of 256 and the validation set was used to determine the validation accuracy. There was slight overfitting, presumably due to the size of the training set. To overcome this problem, a Dropout layer was added and the training was performed with only 21 features,

meaning a dropout of 0.1. This setup resulted in a 72% accuracy. The architecture of the ANN model visualised using VisualKeras library [Gav20] can be viewed in Figure 5.4. The ANN model can be viewed in Snippet 5.2.

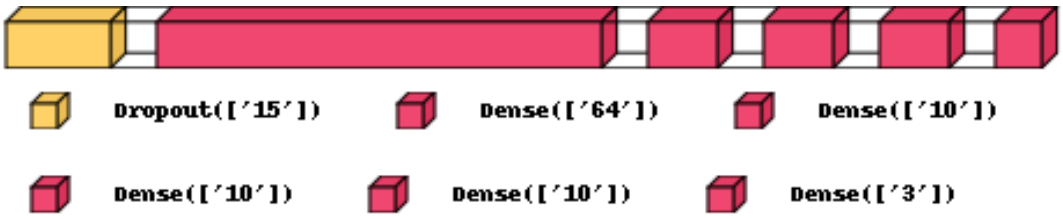


Figure 5.4: The ANN architecture

```
model = Sequential()
model.add(Dropout(0.1))
model.add(Dense(64, input_dim = 13, activation = 'relu'))
model.add(Dense(10, activation = 'relu'))
model.add(Dense(10, activation = 'relu'))
model.add(Dense(10, activation = 'relu'))
model.add(Dense(3, activation = 'softmax'))

model.compile(loss='categorical_crossentropy',
              optimizer='adam',
              metrics=['categorical_accuracy'])
```

Snippet 5.2: Code snippet for the ANN model

5.3.1.6 Convolutional Neural Network

Due to the similarity of data formats, the CNN architecture closely followed the ANN architecture. The first layer was a 0.2 dropout layer to combat overfitting. The second layer is a one-dimensional (temporal) convolutional (Conv1D) layer with 64 neurons, two kernels, and an input size of 5, 6. Five in the input shape represents the rows of data, in this case, 5 seconds of data. The six represents the number of features. Six features are included in the training because of a 0.2 dropout layer. The first dense layer has ten neurons and utilises the reLU activation function. The dense layer is followed by a pooling layer (MaxPooling1D) to downsample the features with a default pool size of 2. The CNN has three more dense layers with a neuron size of 10 each and is finished with the activation layer, utilising the softmax function. The next layer is a

flatten layer. The dataset was used to train the CNN for 20 epochs with a batch size of 256 and the validation set was used to calculate the validation accuracy. The CNN achieved 73% accuracy. The architecture of the CNN model visualised using VisualKeras library [Gav20] can be viewed in Figure 5.5. The CNN model can be viewed in Snippet 5.3.

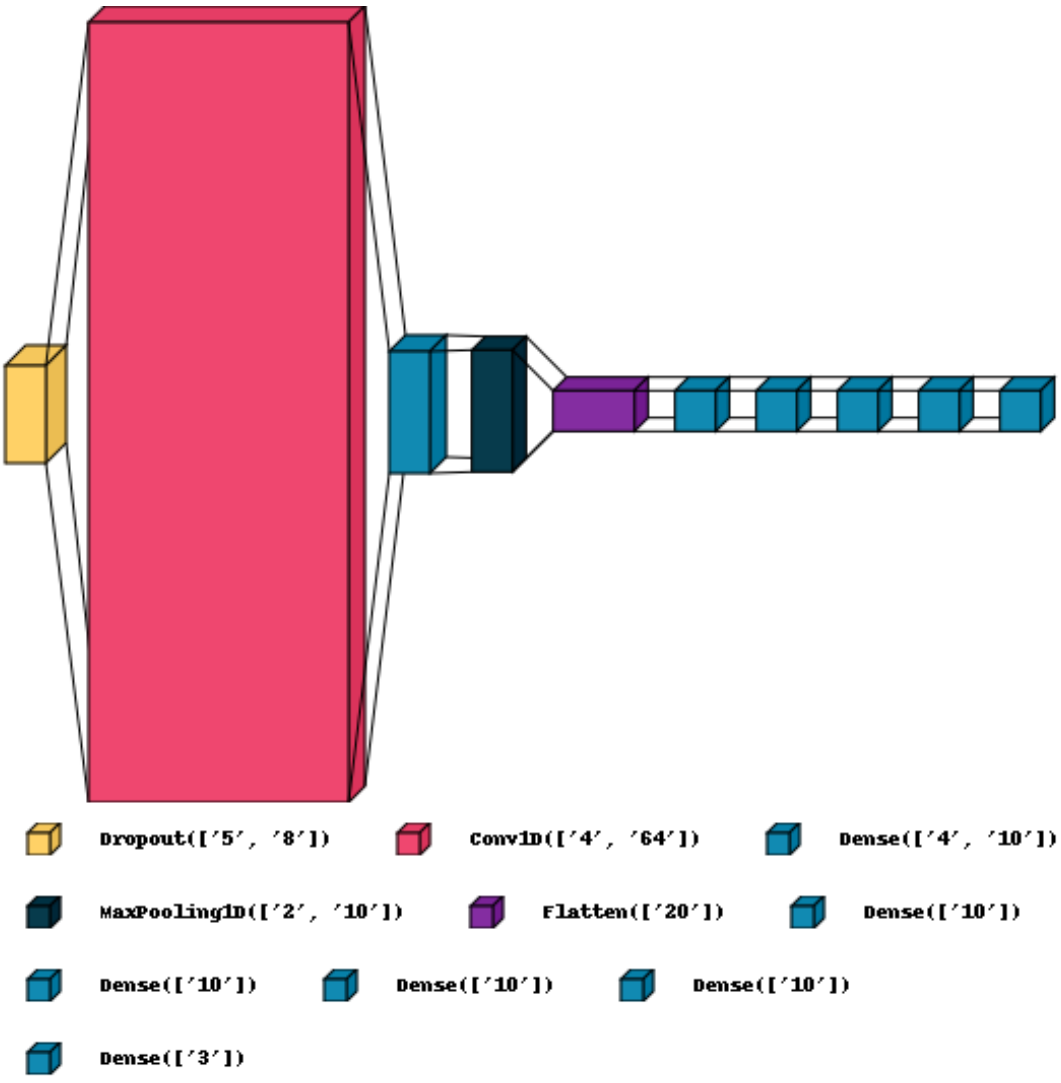


Figure 5.5: The CNN architecture

```
model = Sequential()
model.add(Dropout(0.1))
model.add(Conv1D(64, 2,
    activation='relu',
    input_shape=(5, 7)))
model.add(Dense(10, activation='relu'))
model.add(MaxPooling1D())
```

```
model.add(Flatten())
model.add(Dense(10, activation='relu'))
model.add(Dense(10, activation='relu'))
model.add(Dense(10, activation='relu'))
model.add(Dense(3, activation = 'softmax'))

model.compile(loss='categorical_crossentropy',
              optimizer='adam',
              metrics=['categorical_accuracy'])
```

Snippet 5.3: Code snippet for the CNN model

5.3.2 Leave-One-Out Cross-Validation

Although a simplistic approach was taken to split the dataset into a 20-80 train-test split to compare different models as a means of determining accuracy, this approach has its drawbacks. As the dataset is not very large, how the split is made can significantly impact the resulting accuracy. This technique is sufficient when comparing models but not necessarily a good indicator of the model's actual accuracy.

To achieve results with less bias, all parts of the data must be equally represented. There are several methods to accomplish this. One of them would be k-fold cross-validation. K-fold cross-validation works by selecting k% of the data as validation data, training the model on the rest, saving the accuracy and then repeating k times for each section of the data. The accuracies calculated by each iteration are then averaged to get the final accuracy. A typical application would be 10-fold cross-validation, which trains and evaluates a model 10 times with a 10-90 split.

Similarly, leave-one-out cross-validation (LOOCV) works by splitting one data point or group and using the rest to train the model. This approach is a lot less prone to biases and provides more information by considering every case. However, this can take a very long time in the case of extensive datasets. Since the dataset is not large, LOOCV was deemed the best metric for determining the accuracy.

LOOCV has been applied on both the ANN and the CNN, as they had similar accuracies. In the case of the CNN, an accuracy of 75% was achieved across all

participants. This result is the mean of all the accuracies, but many outliers exist. The median score of 87% represents the accuracy for the majority of the participants. The lowest accuracy was 9%, and the highest was 97%. Results for the ANN are similar, with a mean accuracy of 74%, the highest accuracy of 94% and the lowest of 10%. The CNN performed better by a small margin. Figure 5.6 shows the distribution of accuracies per test participant.

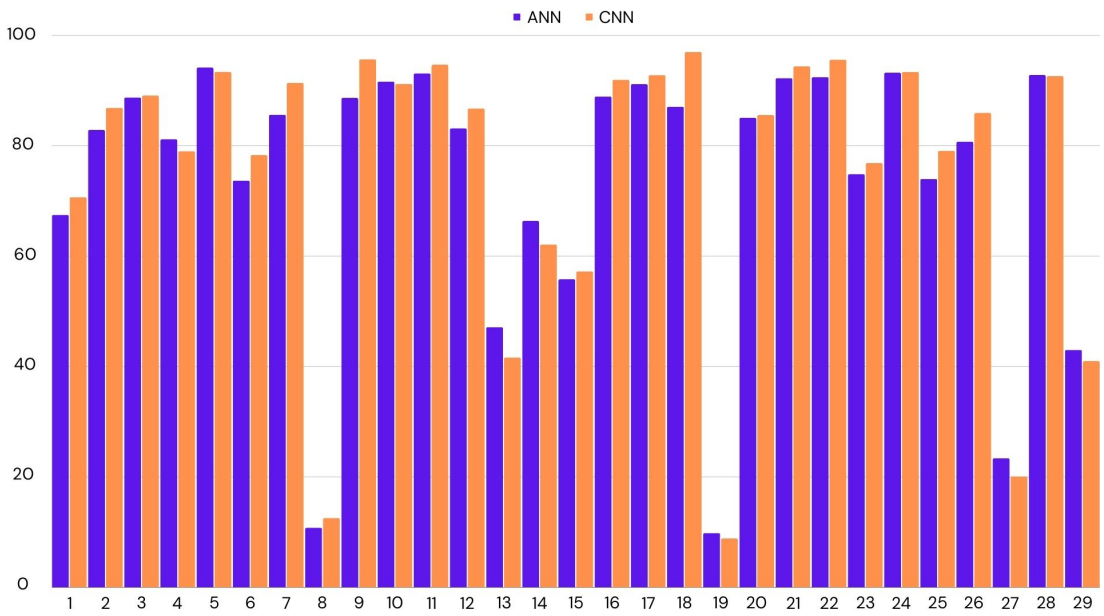


Figure 5.6: Classification accuracy for each participant using leave-one-out cross-validation
ANN: Artificial Neural Network, CNN: Convolutional Neural Network.

To visualise the distribution of class predictions in the model, confusion matrices from each LOOCV iteration were aggregated to create an overall confusion matrix. As shown in Figure 5.7, class predictions were evenly distributed with minimal bias.

5.3.3 Window Selection

Deciding the sliding window size is one of the most crucial steps of time-series classification. Increasing the window size may improve accuracy. However, getting more time-precise output from the classification method is critical for real-time usability. Therefore, the smallest time window was preferred while maintaining accuracy. Time windows of 3 s, 5 s, 10 s and 30 s were tested.

The results were benchmarked using the highest-performing model, the CNN. LOOCV was used for each window size. Table 5.5 shows the accuracies achieved

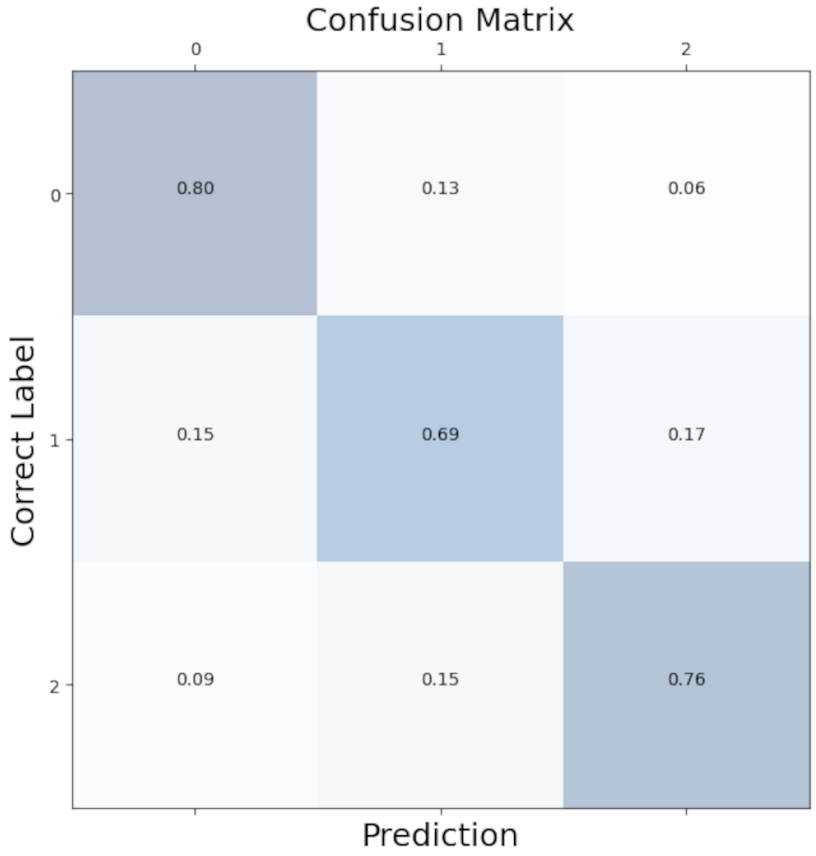


Figure 5.7: Confusion matrix for the Convolutional Neural Network
0: No anxiety, 1: Mild anxiety, 2: Severe anxiety

for each window size. It can be seen that different window sizes performed similarly, providing a reason to use smaller window sizes for more precision in real time. To maintain the best accuracy-time balance, a 5 s time window was used going forward.

Table 5.5: Time windows used and their corresponding classification accuracy

Time Window	**Accuracy**
3 s	74.50%
5 s	74.82%
10 s	74.74%
30 s	76.47%

5.3.4 Metric Comparison

It is vital for inferring each signal's effect on the classification model's accuracy. While the Gini index and ANOVA are insightful, they might not be able to give the whole picture. The best method for measuring the impact of the devices is

to run the models in combinations. In the experiment, three types of signals are utilised: EEG, GSR and PPG. Thus, each combination of the three signals were examined. All the combinations come to a total of seven models, including the combined model. In the interest of time, 10-fold validation was used for these calculations. The results of 10-fold cross-validation were very similar to the results of LOOCV. Aggregated, normalised confusion matrices arising from these calculations can be seen in Figure 5.8.

As visible from the confusion matrices, most of the heavy lifting is done by the GSR signal. The results of ANOVA and Gini feature importances show similar results. This indicates that the EEG and HR signals are not adding value to the model. For more lightweight applications, these signals can be omitted. According to the literature, there is a strong correlation between anxiety and HR, as well as EEG signals. Therefore, the methods employed in this thesis for EEG and HR might not be very reliable. It is also possible that the ground truth method correlates more strongly with the GSR signal in comparison to EEG and HR.

Comparisons between adding the phasic and tonic components of the GSR were examined to see if it is feasible to leave them out for increased performance. The resulting model ended up with 75.03% accuracy, which is very similar to the accuracy in which they are included. This is likely because of the absence of event-related stress, which is more easily recognised by SCR. The SCL signal correlates strongly with the GSR signal. Therefore, it was decided to leave these signals out for the real-time calculations. This amendment improves the optimisation of the system, as CDA calculations can be taxing on the device's performance.

Lastly, all features and the ANOVA-selected features were compared. The accuracy of the model, which was trained in all features, was 74.71%. The accuracy of the model trained on the F value selected features was 75.38%. This shows that the feature selection method resulted in a slight increase in accuracy.

5.4 Discussion & Limitations

The protocol followed in this chapter goes through the steps of data acquisition, data preparation, feature extraction and anxiety classification for use in real-time VR applications. It has many similarities to methods used in the literature for anxiety classification using biosensors. The methodology follows

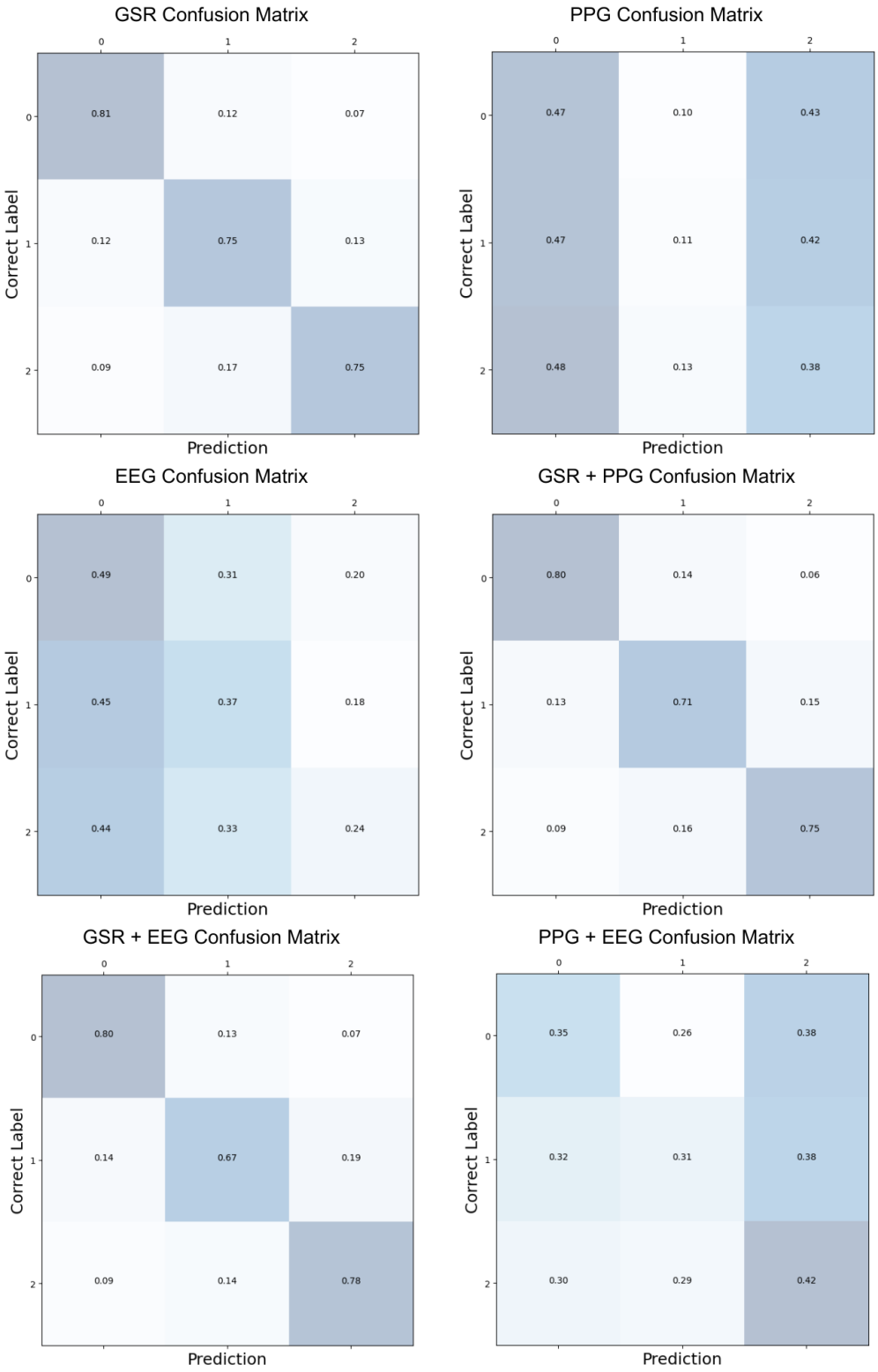


Figure 5.8: Confusion matrices per signal combination

past examples of using well-established cognitive tasks to label data. It uses established metrics for classification as features: HR, GSR, alpha, beta and gamma EEG bands.

Explainability was prioritised at the setup of this analysis; hoping to get accurate results by establishing a rule-based system using a decision tree as a starting point. However, due to the complex nature of the interactions between different physiological measures, as well as the interpersonal differences in baseline physiological values, it was not possible to achieve high accuracy using simple models. The accuracy achieved by neural networks outperformed those by decision trees by more than 15%, even after fine-tuning the trees. Therefore, a decision was made to go forward with CNN.

At the setup of these experiments, utmost attention was given to rule-based systems that can be used to classify different states of emotion and anxiety. The reason for this is the simplicity, explainability, and light-weightedness of such algorithms. Although significant effort was exerted in this direction, it was not possible to develop a rule-based system with sufficient accuracy. Therefore, a black-box approach was adopted.

Convolutional neural networks are a common approach to classification for time-series data, and they have the advantage of reducing data preparation time by not needing rolling windows. Many examples in the literature use simple convolutional network structures to analyse biosensor data for outputs such as emotion recognition and stress recognition [HLL+19, GMSSMFL22, LSZM21].

The current work, however, differs in a few accounts. Firstly, for real-time use, smaller windows and different data-cleaning methods were employed. For example, applying Blackman filters per second can make extracting meaningful data more challenging than applying such filters over a long period. An instrumental metric, HRV, was also deemed impractical due to the short window lengths.

Additionally, the current work has to deal with additional noise and unreliable signals due to VR electrical interference and excessive arm movements. Although biosensors produce much more accurate results when stationary, allowing arm movements is crucial for controls in VR. The dataset was collected in a VR state of the art emotional Stroop task environment to 1) guarantee that what is being measured is indeed mental anxiety and not task load and 2) the data that is being collected is similar, therefore applicable to the use case of VR

therapy and 3) while it is applicable to the scenario, it also applies to different types of therapy by being a generalisable task.

The model is evaluated based on classification accuracy. Based on the leave-one-out cross-validation results, the model successfully predicts the correct label 75% of the time. However, it must be noted that based on the LOOCV results, there can be vast variations based on the specific participant. While the median score was higher than 90%, the accuracy suffered levels around only 10% for two participants.

The model correctly predicted the "no anxiety" condition 80% of the time and the "high anxiety" condition 76% of the time. The model had difficulty with the "mild anxiety" condition. 15% of the mild conditions were incorrectly labelled as no anxiety, and 17% were incorrectly labelled as high anxiety. 69% of the mild condition was correctly predicted. This disparity reflects the difficulty with clear-cut separations of fluid emotional anxiety levels.

The outliers in the model were examined for any possible explanations. Looking through the demographic information of these two participants, only one participant was an outlier in a category. They were one of the two people diagnosed with dyslexia in the dataset. Their dyslexia could be why they are an outlier; however, there is insufficient information to draw any conclusions. It must be noted that this method of anxiety classification might not be ideal for participants with dyslexia.

It must be noted that prior to machine learning analysis, it was planned that the neutral level of eStroop would be used for baseline analysis in practical applications. However, using real-time analysis simulations seems implausible due to not representing a wide range of anxiety. The dataset's standard deviation and mean must be known for normalisation reasons. However, the mean is a lot lower in the neutral condition. Therefore, the normalisation and analysis are only accurate when the complete eStroop task is administered.

5.4.1 Comparison to Similar Work

The literature that focuses on real-time anxiety detection for VR is limited due to complications with real-time signal processing, as well as the limited research on VR. Some studies employed longer time windows, making it unsuitable for real-time predictions [BFF21, SAE+15, ZWLH16]. Many studies with shorter time windows have limited participant numbers or are preliminary studies

[BMM+20, CHO+17, HCOL17, PPM+20]. Generalisation can be challenging in such studies. A summary of the comparison of the results with previous studies can be viewed in Table 5.6.

It can be seen that previous studies with three output levels and short time windows (3s-30s) range between 67% and 96% accuracy. The closest condition to the current study was one conducted by Wu et al. [WCL+10], using the Stroop Colour-Word Task (SCWT) for stress inducement and classifying between 3 output levels. They also had 19 participants, closer to the current study than the others. They achieved 84% accuracy using an SVM. However, it must be noted that in the case of traditional SCWT, it is difficult to differentiate mental load from emotional anxiety.

It must be noted that all of the studies in the literature use different methods for ground truth and labelling anxiety, which makes direct comparisons challenging. The current study differs from previous studies because a new way of measuring ground truth is introduced, making it implausible to do a direct comparison. The current study also avoids using computationally heavy processing or processing that requires full datasets so that it can be applied in real time.

5.4.2 Limitations

There are some limitations due to the representation of the dataset sample. The dataset includes 29 participants recruited from a large university campus, which may not accurately reflect the general population. Additionally, the method of separating anxiety based on emotional words in the current study might not reflect the full breadth of the anxiety spectrum, making it challenging to separate very high levels of anxiety from high-level anxiety.

Furthermore, this review only addresses methods for decreasing noise in real-time using filters and does not propose any physical methods to account for noise in signals. Validated APIs and software by publishers endorsed by devices are used, and the experiment makes no contribution to data acquisition methods.

Raw signals are used for GSR, and PPG is converted to HR in real time. Additionally, band information is extracted from raw EEG signals in real time. For ideal results, developing an algorithm that can separate phasic and tonic skin conductivity signals within a five-second timeframe would be necessary. Similarly, developing an algorithm that can reliably measure HRV from the raw

Table 5.6: Comparison to similar studies

Ref	Sample size	Features	Ground truth	Highest accuracy model	Number of outputs	Accuracy
[BFF21]	8	HRV (ECG)	VR Content	LSTM	2, 3, 4	90.5% (2-Level), 67.5% (3-Level), 58.8% (4-Level), 30s
[BMM+20]	4	PPG, EDA, EEG	SUDS	ANN	11, 4, 2	78.3% (2-Level), 38.8% (4-Level), 26.5% (11-Level)
[CHO+17]	12	PPG, EDA, SKT	VR Content, STAI-Y1	K-ELM	3	96.3%
[HCOL17]	6	ECG, PPG	VR Content	LDA	3	79%
[KSZ21]	28	EEG, EMDR	SCWT	MLP	2	96.42%
[PPM+20]	7	EDA, PPG	SUDS	Regression	2, 3	94.3% (2-Level), 92.4% (3-level)
[TCT+15]	20	ECG, RESP	STAI-Y1	Neuro-fuzzy	4	83%
[vDML19]	30	BVP, EDA, SKT	VR Content, SUDS	SVM	4	86.3%
[WCL+10]	19	EDA, RESP, ECG, EEG	SCWT	SVM	3	84%
Current Work	29	EDA, EEG, PPG	eVRST	CNN	3	75.38%

Note: ANN: Artificial Neural Network, BVP: Blood Volume Pressure, CNN: Convolutional Neural Network, ECG: Electrocardiogram, EDA: Electrodermal Activity, EEG: Electroencephalogram, EMDR: Eye Movement Desensitisation and Reprocessing, EMG: Electromyogram, HRV: Heart rate variability, K-ELM: Kernel Extreme Learning Machine, LDA: Linear Discriminant Analysis, LSTM: Long Short Term Memory, MLP: Multi-layer Perceptron, PPG: Photoplethysmogram, RESP: respiration, SKT: Skin temperature, SCWT: Stroop Colour-Word Task, SUDS: Subjective Units of Distress Scale, STAI: State-Trait Anxiety Inventory, SVM: Support Vector Machine

PPG signal in this short time would be optimal.

5.4.3 Recommendations

As shown in the results of the machine learning analysis, skin conductivity is doing the heavy lifting in all of the algorithms. This is due to the close correlation between arousal and skin conductivity. However, heart rate and brain activity are important measures closely related to the emotional state. Brain activity is one of the most reliable methods of measuring the direction of arousal, positive vs negative. Therefore, it is imperative to research a more ergonomically feasible way of measuring brain activity than the current solution.

Additionally, the current method of measuring ground truth is focused on generalisation and emotional anxiety. However, it is likely fine-tuned to a limited range of anxiety. It might be beneficial to research ways to measure a bigger anxiety range and test it in different scenarios. The resulting algorithm for machine learning will be used as part of a decision-support system and integrated into exposure therapy in Unity.

5.5 Conclusion

This chapter showcases the real-time anxiety classification system for seamless and objective feedback in VRET. A 75% accuracy between three output classes was achieved, showing that it is possible to predict anxiety in real-time with no access to signal processing tools that rely on complete datasets.

The real-time aspect of the algorithm is emphasised by keeping the sliding windows down at only 5 seconds. Keeping the time window small increases the reliability of the results by showing more recent information. Some algorithms may claim real-time results, but the algorithms they use employ time windows of around a minute.

It must be duly noted that these results show that it can be challenging to apply the classification to some users with the same accuracy, as evidenced by the LOOCV results. For some participants, the classification accuracy was as low as 9%. It is essential that factors that cause such outliers are investigated for the general applicability of these results.

The created model will be integrated into the Unity environment for real-time classification. By doing so, it will be possible to accomplish seamless decision

support for the therapist. Creating a generalisable system for anxiety detection is the goal. However, thus far, the model has only tested the system in the eVRST system. Therefore, the first step after integrating into the VR framework will be to validate it within the exposure environment.

There is some evidence of anxiety classification within VRET providing optimal results. However, no research exists to see if such systems with seamless feedback improve therapy outcomes. It is crucial to conduct experiments with actual patients to see the effect of such decision-support frameworks.

This chapter made a detailed analysis of the changes of negative arousal between different conditions. The accuracy of the developed algorithm surpassed 75%, showing that it will be effective in differentiating between different anxiety conditions. The following chapter will build on this algorithm, and create a framework that can classify a patient's anxiety in real-time during a VRET experience.

Chapter 6

Virtual Reality Exposure Therapy with Seamless Feedback: A Prototype

There have been extensive applications for VRET in recent literature, with VR being a focus point in mental health research and computer science research. The applications vary between different mental health conditions, ranging from PTSD to anxiety disorders. There are examples of specific phobias such as public speaking phobia, fear of heights, and fear of flying.

Some environments integrate seamless objective feedback into a VR environment, such as the work of Salkevicius et al. [vDML19], showing that the concept is possible. The current chapter focusses on the robustness and generalisability of the system, incorporating seamless feedback into many different phobia scenarios.

This chapter details the development of a framework that can aid users in safely experiencing their feared stimulus. It provides an interface for a mental health professional to review user status and allows a mental health professional to adjust exposure levels to the feared stimulus. The framework also offers both AI predictions and raw physiological data in the information panel, and it is designed to be easily scaled and transferred for different uses. A bird's eye view of the integration can be viewed in figure 6.1.

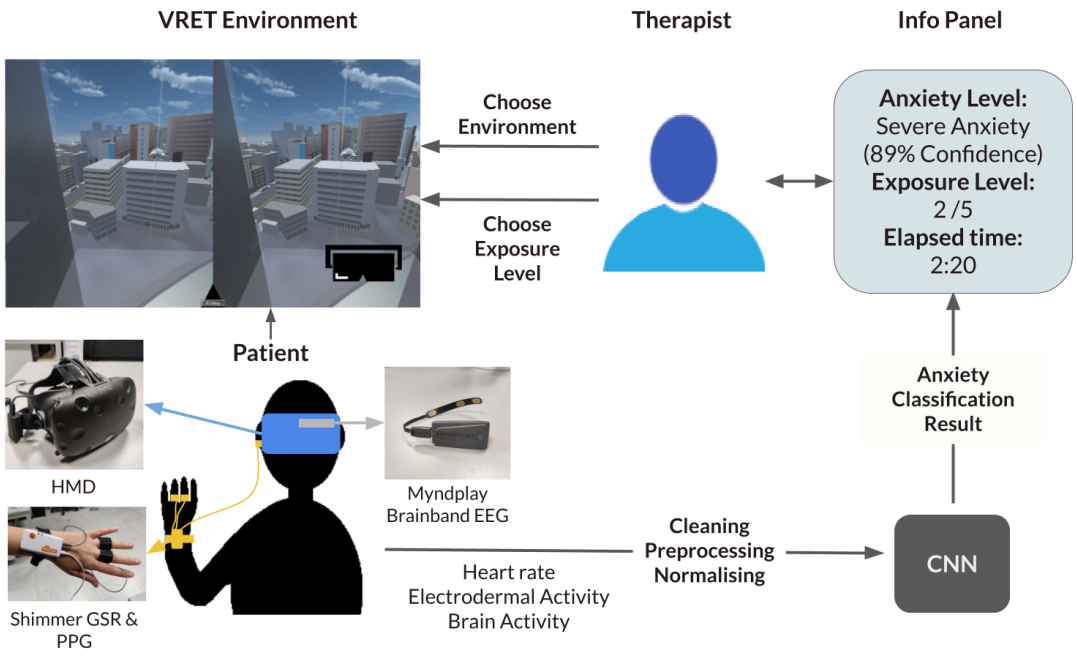


Figure 6.1: Framework Diagram. CNN: Convolutional Neural Network, EEG: Electroencephalogram GSR: Galvanic Skin Response, HMD: Head-mounted Display, PPG: Photoplethysmogram.

6.1 Background

Virtual Reality Exposure Therapy is now widely accepted as an effective method of therapy for anxiety disorders [CSLC+19a, EM21]. The tools used for administering such therapy sessions vary between applications. There are no generally accepted environments for administering such therapy, with many researchers and healthcare professionals using their own version. Additionally, there are no generally accepted user interface and user experience values for VRET applications. This makes it difficult to compare different applications and creates barriers to effectively administering the treatment.

Different anxiety disorders such as SAD, specific phobias, and GAD each have a range of applications for their treatment. Some of the most notable examples of this include studies by Meyerbröker et al. [MMKE13], Krijn et al. [KEÓ+07], Kampmann et al. [KEH+16], and Bouchard et al. [BDR+17]. These environments were created to administer therapy and measure efficacy compared to waitlist and in vivo controls.

There is a current lack of standardised methods of feedback and a lack of unity between methods of administration in between applications of VRET. A system

that combines such therapy methods and allows for personalisation and context adjustment is vital.

The integration of seamless feedback in VRET systems is crucial for enhancing therapeutic outcomes. Real-time feedback on physiological indicators allows for timely adjustments to the exposure scenarios and maintains an optimal level of exposure for the patient. This adaptive approach not only helps prevent overwhelming the patient but also ensures that the exposure remains effective by staying within the therapeutic levels of anxiety.

While there have been many recent advances, the variety of research has led to difficulties due to different systems using varying physiological and behavioural measurements, ground truth methods, and therapeutic protocols. This heterogeneity makes it difficult to standardise and scale such applications. The variability in target conditions requires a flexible solution that can be easily customised and scaled according to the specific needs of different populations.

Moreover, the user experience (UX) in VRET systems varies significantly, increasing the difficulty of therapists adapting such tools into their practice. An intuitive and comprehensive user interface (UI) system is vital for the seamless transition from current therapy methods into VRET with physiological feedback. Ensuring that therapists can easily interpret the data and make real-time adjustments without extensive training is critical for the practical implementation of VRET systems.

Ultimately, the goal is to develop a VRET framework that is not only effective but also accessible and easy to use for both patients and therapists. Such a framework would facilitate the personalisation of exposure therapy, improve patient outcomes, and promote the standardisation of VRET practices across different therapeutic settings. This chapter will explore the design and development of such a framework, emphasising the importance of seamless feedback integration and the potential for generalisability across various anxiety disorders.

6.1.1 Related Work

Gaggioli et al. [GCS+14] proposed a decision-support system (DSS) for automatic classification of stress in VR. They used ECG, RESP, EEG and body movements as predictors and used self-report measurements from patients and therapists to establish ground truth. They detected stress events and used a

web interface to deliver this information via a graph. For the automation of feedback, they used a combination of fuzzy logic and ANN inference.

Salkevicius et al. [vDML19] created a framework that classifies anxiety in real-time using four output levels (low anxiety, mild anxiety, moderate anxiety and high anxiety) with an accuracy of 80%. They used GSR, BVP and SKT for the calculations and a window size of 20s. Their cloud-based application contains a dashboard for the psychologist and a VR environment for the patient. It is compatible with multiple exposure scenarios, which can be swapped through the dashboard. The available options are public speaking anxiety, fear of heights, and panic disorder.

Rahman et al. [RBS+22] took a different approach by distancing the therapist from the treatment and making a self-guided application. Using HRV and EEG data, they predicted anxiety with an accuracy of 87.5% in time windows of 5 s. Using this feedback, the user is able to change the exposure level in a public speaking anxiety scenario.

6.2 Design

The framework design focuses on usability, non-interference, and visual attractiveness. Every step of the way is optimised for time and effort in order to decrease the dropout rates in therapy. The design is intuitive, efficient, and user-friendly.

6.2.1 Virtual Reality Exposure Environments

The framework has been structured to support as many exposure environments as possible. While the framework has only been tested with two exposure environments for proof of concept, it can be integrated into any scenario with minimal adjustment.

6.2.1.1 Acrophobia

Acrophobia, or the fear of heights, is a common phobia that can severely impact a person's daily life. VRET provides a safe and controlled environment where patients can confront and desensitise from their fears. A glass elevator scenario was developed because it is common in daily life and its natural graduation between heights, which is crucial for exposure therapy (Figure 6.2).

To administer the exposure effect in the framework prototype, an acrophobia environment which was developed and validated by the authors during their master's degree [MTM21b] was adopted. This VR environment places the participant in a glass elevator overlooking a large city. In the original experiment, the participants spent a predetermined amount of time at each height level before deciding to go higher by pressing a button. This environment was originally used as part of a study investigating methods of visual respiratory feedback [MTM21b]. As a side analysis, it was found that the users' arousal increases as the height increases in the virtual environment, marked by a strong correlation between height and GSR. This correlation confirmed the predetermined goal of a similar effect of height to that experienced in real life.

Some adjustments were made to make this pre-developed acrophobia environment more suitable for the current framework. Firstly, the self-controlled buttons and predetermined wait periods were removed and replaced with therapist-controlled UI. The height levels were modified to correspond with five separate exposure environments, with equal height distance between each consequent exposure level. The heights in each exposure level are as follows:

1. Level 1: 10 m
2. Level 2: 20 m
3. Level 3: 30 m
4. Level 4: 40 m
5. Level 5: 50 m

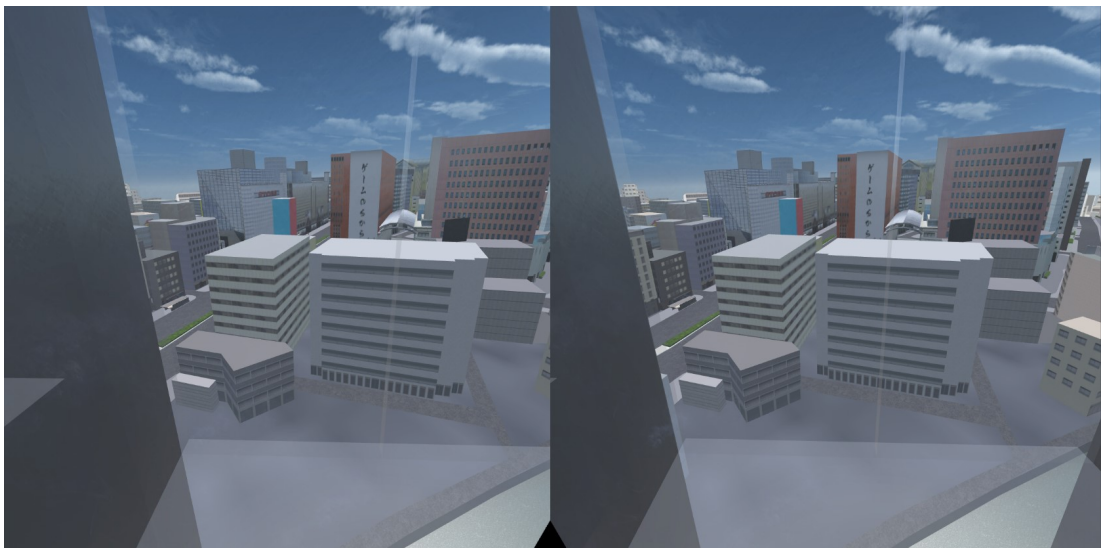


Figure 6.2: Perspective of the user looking out of the glass elevator, both eyes rendered

The environment aims to be fully immersive, equipped with realistic haptics, spatial sound and movement. Thanks to the use of a sonic transducer, the user can feel the vibrations of the elevators as they ascend or descend. Based on a prior study, the mean motion sickness of the eleven participants out of a Likert-scale of 5 was only 1.45, meaning they strongly disagreed with having felt motion sickness. Additionally, reports indicate that the haptic feedback increased the immersion and presence in the environment [MTM21b].

The elevator script in this environment uses the IExposure framework with the required function of ChangeExposure(int level). This function is required for each environment in the framework. It is called when the therapist transitions the user between different exposure levels. It enables standardisation between separate exposure environments, making controls smoother.

6.2.1.2 Claustrophobia - Proof of concept

This environment was created to demonstrate the ability to transfer the framework from one environment to another. As it is only for demonstration purposes, there is no validation of its effect on claustrophobia. It is a minimal example.

It uses the IExposure framework and, therefore, has the required function of ChangeExposure with five levels. Each of the five levels of exposure is a room with decreasing size, decreasing light, increasing clutter and decreasing windows. Although they were created with common triggers for claustrophobia using prior research, this environment has not currently been validated to produce the desired effect. It has only been used for development and testing purposes.

The five levels of exposure rooms are structured as follows:

1. Level 1: Living Room - A large, well-lit living room with ample space, minimal clutter, and two ceiling-length windows.
2. Level 2: Big bedroom - A slightly smaller room with moderate lighting, some clutter, and one medium-sized window.
3. Level 3: Small bedroom - A medium-sized single bedroom with dimmer lighting, a single bed, a bookshelf and a desk, and one small window.
4. Level 4: Box office - A small room with dim lighting, a desk, a cabinet, and no windows.
5. Level 5: Bathroom - The smallest room with dim lighting, no windows and barely enough space to stand.

Each room can be viewed in the screen capture Figure 6.3.

The ChangeExposure(int level) function is central to this environment's operation. It changes the user's environment by moving the HMD camera to a different room, each corresponding to a different exposure level. This transition between different rooms is vital for maintaining engagement and appropriate arousal levels while claustrophobic stimulus gradually increases.

6.2.1.3 Nature Scene

The same nature scene used in the eVRST application was also used here for similar reasons. This is a way for the patient to relax before they are ready to



Figure 6.3: From top to bottom: Living room, big bedroom, small bedroom, box office and bathroom

come out of the VR environment if their anxiety increases. It is accompanied by instructions for breathing exercises.

6.2.2 Graphics

The elevator scene was built using free assets from the Unity asset store to save time. While choosing the assets, priority was given to performance and realism. The scene overlooks Fukuoka, a city in the Fukuoka prefecture of Japan. The elevator is stylised, and a glass-looking material was created for extra visibility. This material was made transparent and metallic to make it reflective, and some scratches were normal-mapped onto it to make it look more realistic. Soft shadows were added to avoid the unnatural look of default transparent shadows on Unity.

The elevator frame was matched to the height of the building by which it stands. The height is approximately 60 meters. The elevator's speed was roughly matched to the speed of the glass elevators in the University College Cork Western Gateway Building by measuring the time it takes to traverse the same height. This was done to increase realism and make integrating audio recorded in the same location easier.

The scene includes three light sources: ambient light modelled to imitate the evening sun, a dimmed spotlight modelled to imitate the overhead light in an elevator, and a point light over the river that flows in front of the elevator to enhance the look of the water shader.

The Claustrophobia environment was built using ProBuilder and free furniture assets from the Unity asset store. The living room contains a lamp, a sofa, a TV set, a TV cabinet, a coffee table, a dinner table, six chairs and two large windows. A point light is attached to the lamp. For this scene, window shaders were not used. Instead, the skybox attached to the camera is visible through the windows.

The big bedroom contains one double bed, sofa, dresser, lamp, end tables, and window. The small bedroom contains one single bed, one desk, one chair, one laptop, one bookshelf, and one window. The box office contains one lamp, an office cabinet, a desk, a chair, and a computer. The bathroom contains one toilet and one sink.

6.2.3 Audio and Haptics

Spatial audio is vital for immersion in VR environments. Ambient nature sounds are ambisonic and are implemented using Google Resonance Unity Package. The mixer for spatial sound from the resonance plugin was used to do this, and the audio file was tagged as ambisonic.

In the acrophobia scene, several audio files are integrated concurrently. The ambient audio was spatialised around 1m in height. Ambisonic audio was not used due to the vertical distance of the elevator from the cityscape as the patient ascends. The elevator music was placed on a corner of the elevator, modelled to imitate a speaker.

The elevator audio was recorded using one of the glass elevators in the University College Cork Western Gateway Building. Sennheiser AMBEO first-order ambisonic microphone was used to record the audio. The microphone was placed in the middle of the elevator at approximately 175cm, facing the elevator door.

Clark Synthesis TST329 Gold Transducer was used for vibrotactile feedback to imitate the elevator's movements. The haptic speaker was placed under the floor tiles in the laboratory. By integrating haptic feedback, increased immersion was sought after. To synchronise the vibrations of the elevator with the audio in the experience, high-frequency waves were filtered from the original recording and the low-frequency audio was played through the transducer.

The audio file for use with haptic speakers was created in Reaper by applying a low-pass cut at 100 Hz frequency to the ambisonics file and downmixing it to one channel. This frequency was chosen because human skin is sensitive to vibrations of up to 160 Hz [YMW15].

6.3 Seamless Feedback Integration

A Flask Python server was to implement the predictions and make post requests to the server from Unity to update it. This method was chosen after trying both in-Unity and server options.

6.3.1 Unity solution

The built-in machine learning tools in Unity are neither extensive nor regularly updated. There are extensive uses for reinforcement learning using agents to

train non-player character behaviour using the plugin ML-Agents. However, when it comes to applying classification, the integration is suboptimal. There are packages such as Barracuda that apply neural networks to Unity scripting. However, the primary purpose of the plugin is to be used with ML-agents, which is not relevant in the current scenario. The package requires many additional packages to be imported, bloating the application. Barracuda only supports inference on pre-trained models and does not support training with new data. Machine learning models must first be converted into the ONNX file format to use pre-trained models with Barracuda. There are limits to the Keras operators that can be used in Barracuda. The operators used in the model, however, are all supported. The latest version of Barracuda was downloaded and imported. The installation comes prepackaged with a dll of Google.protobuf. This requirement clashes with the Shimmer Library, which requires a specific version of Google.protobuf. Having dlls with the same name is not supported in Unity. Therefore, it was required to change the name of the dll and aliases and change the assembly of each package to strongly name a specific version of the dll. This workaround is not an optimal, permanent solution because it might not persist through Barracuda version updates.

The Barracuda ModelLoader method is used to instantiate the pre-trained CNN model from Keras. It was first necessary to convert the saved model to an ONNX format using the tf2onnx via the command line. The model was saved in .pb format and then the tf2onnx library was installed to do the conversion. However, this library does not support all operators embedded into Keras and must be used cautiously. In the case of the VRET application, all of the operators used in the model were supported, and the conversion was successful. Barracuda uses an Interface called iWorker to handle machine learning inference. A worker is created from the WorkerFactory using a backend and loaded model. As the inference is simple and the VRET Unity application mainly uses the GPU, it was decided that CSharpBurst was the optimal backend. With a simple model and data, GPU communication can take more time than CPU execution. CSharpBurst is a highly efficient and parallelised CPU code compilation using Burst, a Unity compiler optimised for large language models. The CPU code is jobified in this backend option to work well with Burst. A job is a struct that lets the code run parallel in a multithread, increasing performance and efficiency. After creating the worker, tensorising the data is required. Tensorising is done by creating a new tensor with the correct batch size, height, width and channels. It must be noted that the shape is

different from that of Keras, where channel size comes first. As the application uses single interference, the shape is (1, 5, 8, 0). The worker execute method is then used to run the inference on tensorised input. The PeekOutput method is used to receive the outputs. The PeekOutput method is preferable to Fetch because it disposes of the tensor after the next execution, providing efficiency in garbage management. The output is then displayed in the application.

6.3.2 Server solution

The machine learning model that was created is saved in a .h5 file. The .h5 file is loaded onto the Flask application using the load model method of TensorFlow. The Flask webserver is run locally on a port. The predict function uses the JSON component of the request. The information is then normalised using the eVRST mean and standard deviation and structured into a 2D NumpyArray in the shape of (None, 5, 8) to match the data structure of the CNN model. The model runs the prediction and returns the predicted label and the confidence level. To receive the output, the predict() method of the model is called. The predict() method returns a vector of probability scores for each output level. The argmax function is used to pick the highest value. The corresponding label to argmax index is returned as a string, and the value is reformatted and returned as an integer.

On the Unity 3D side, UnityWebRequest is used to send the data in JSON format. In order to do so, the JSON is first encoded into bytes. A new UploadHandlerRaw and DownloadHandler are requested to handle the data. The request header is set to Content-Type, application/json. If the request is unsuccessful, the error code is returned using downloadHandler.text. If the request is successful, then the prediction is pasted on the UI information panel.

6.3.2.1 Containerisation

A container is an environment that bundles the code with all the files and libraries so the application can efficiently run on any platform. It avoids needing specific versions of libraries, software, and operating systems to run the application. Containers are lightweight and are portable. They work similarly to a virtual machine but run straight from its kernel and are smaller and faster to start. They prevent the errors arising from version mismatches and remove the need to share requirements and libraries when transporting software. It is also easier to manage through built-in deployment, upgrade, and rollback

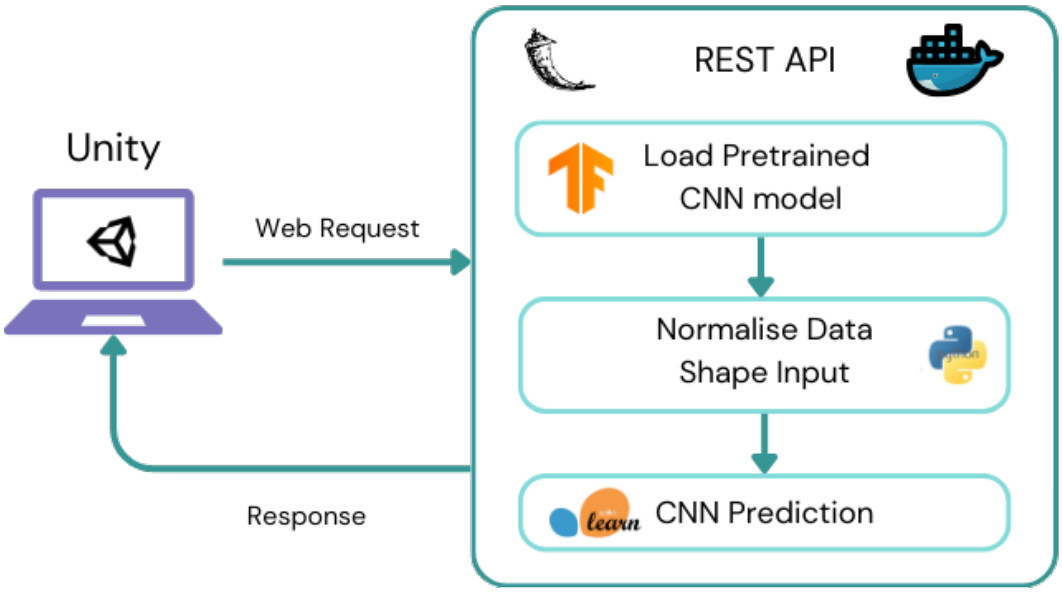


Figure 6.4: Machine learning communication diagram

processes. Sharing work and collaborating is much easier after overgoing the requirements and platform differences. It makes it very smooth to deploy the application onto the cloud.

Docker is the industry standard for containerisation. Docker makes it easy and fast to containerise any application with minimal effort. A dockerfile is a text file that has a set of CLI instructions for creating a docker image. This often includes instructions on installing the packages and libraries specified in another text file. Docker operations work the same regardless of infrastructure. A docker image contains all the necessary information and dependencies to run the application from the docker container. Every time a new image is created, the older ones are saved as layers, simplifying version control.

In this project, Docker is used to build and deploy a container. The dockerfile includes the OS architecture Python 3.9. It specifies the Flask application and the host and copies the necessary code files and pip instructions to install the requirements. Lastly, it exposes the port and uses the Python command to run the application. The requirements are TensorFlow and Flask. Numpy comes as a prerequisite to TensorFlow. Docker compose was used to start and stop the services. The compose.yaml file includes the ports and version of Python.

Docker hub is a registry for open-source containers, making sharing docker images easier. This way, it is possible to pull the repository and deploy it

quickly and without error from anywhere in the world.

6.3.3 Comparison and Decision

Integrating machine learning into Unity is not a straightforward process. Some functionality is compromised, and certain libraries have version mismatch problems. It is also not scalable or maintained as often as Python libraries. When Unity gets updated, the processes for changing dlls must be applied again. Additionally, C# does not have as many built-in tools for data manipulation as NumPy for Python. It also lacks portability to the server solution's different types of applications. It has the advantage of being prepackaged; for built solutions, it would come in integrated and would not require a server to be run.

The server solution, which uses Python, Flask, and Docker, has many advantages over Unity integration with Barracuda. It is transferable, portable, fast, quickly run and lightweight. Thanks to the Docker Hub, it can be downloaded as a Docker image from any machine and will always behave the same way. It can be updated separately without changing the Unity application and Unity packages. It fully supports all Keras operators and NumPy operators. It is effortless to set up directly from Keras after training. In the future, where fine-tuning might be applied based on the user, it is easy to integrate it directly.

All in all, the benefits of the server solution far outweigh the benefits of Barracuda. The Flask server is deemed to be the optimal solution for the VRET framework. The Flask server can either be run locally or on a cloud for further efficiency and performance.

6.3.4 Real-Time Processing of Data

The machine-learning model involves minimal data processing and primarily uses the APIs by with Shimmer and NeuroSky. The PPG and GSR signals are filtered in real time using Blackman FIR filters. Frequencies above 5 Hz and below 0.5Hz were filtered from the PPG signal. Frequencies above 5 Hz were filtered from the GSR signal. Real-time application of the filter is not as efficient as the ones applied to the whole signal.

After processing band rates, heart rate, and skin conductivity, outliers are removed by measuring values that are three standard deviations apart from the mean. This is done in real-time over three minutes to ensure that it is not just a spike in emotions but noise in the wearable devices.

To input data to the CNN, the data first has to be normalised. This is especially important because physiological signals can be different between people. To do this, the information from the eStroop environment is used. The mean value of the signal in the eStroop part is removed from the real-time signals and then divided by the standard deviation.

6.4 UI

The UI is created using the built-in Unity tool. The UI is intended for the therapist's use only, so it is not rendered in the 3D VR space. Instead, it is attached to the preview visible on the computer and only viewable by someone not using the HMD. The built-in Unity tool provides some useful premade objects such as text, buttons, options, and images. The buttons are interactable using a mouse, and custom functions can be attached to them. They also have some default styling options with outlines around and text in the middle.

The UI is comprised of three different components: option menu, main menu and graphs. The main menu is enabled by default and is intended to stay on throughout the experience. It includes an information panel which relays the current ML anxiety, average anxiety through the exposure level, time spent in therapy, time spent in current anxiety level and the current exposure type and level. Underneath the dropdown menus are four buttons: apply, options, nature and advanced. The apply button applies the selected settings to the environment. The nature menu takes the patient into the Nature environment to relax. The options button enables the option menu which includes volume control. The advanced button enables and disables the advanced graphs. The main menu UI with the advanced graphs enabled can be seen in Figure 6.5.

6.4.1 Data Visualisation

Advanced options for viewing raw data are available for therapists who feel comfortable with it. This helps supplement and explain the information given by the machine learning algorithm. Visualisations are in the form of line graphs derived from raw EEG, GSR, PPG and HR derived from PPG. This helps see trends over time and enhances the progress of decision-making. It can be visible, but it can also be hidden for those who do not want to use it.

The graphs were created using the XChart plugin. A transparent background was used for the line chart class of XChart. In all of the graphs, the X axes

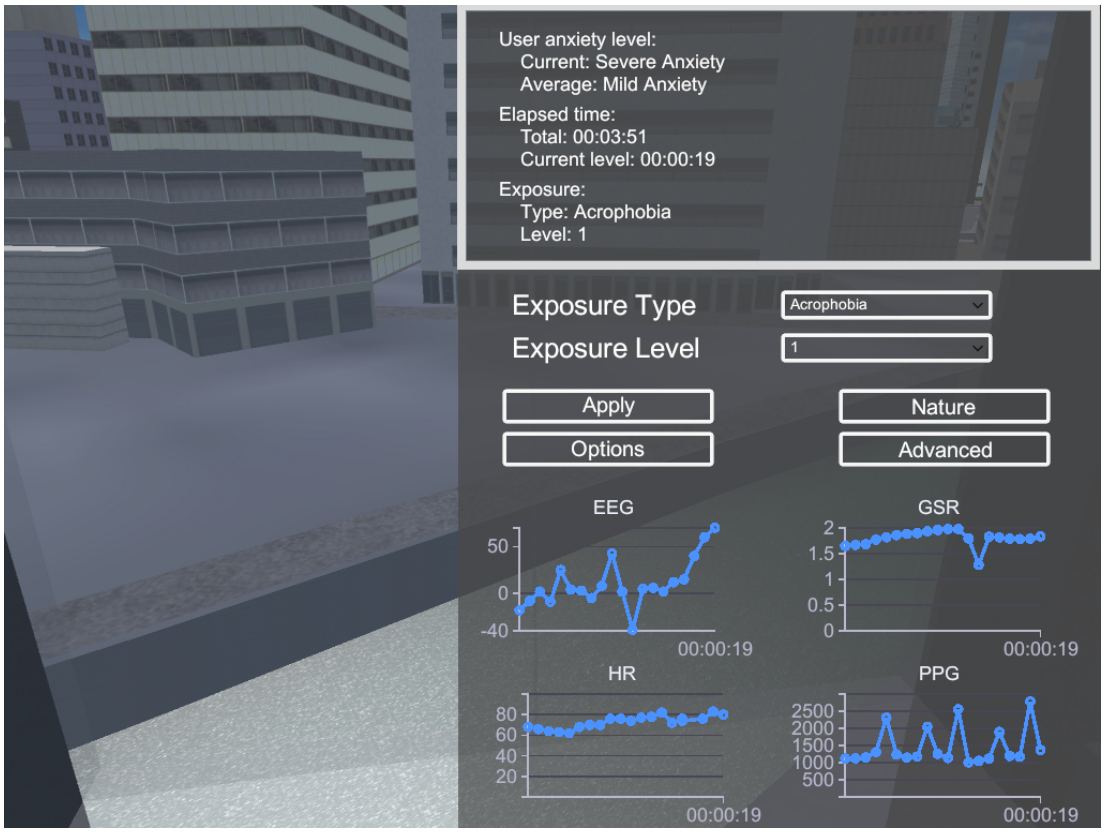


Figure 6.5: UI

represent time in seconds. As the data is downsampled to 1 Hz, the graphs also update once a second. The maximum number of data points is set to 60s for better visibility. This means only the last minute of data is available. Each data point can be viewed by hovering over it with the mouse. The data is visualised only when the advanced panel is activated. Therefore, the data points update over time and do not immediately display the last 60 seconds of data points.

6.4.2 Level Manager

The UI manager handles changes applied by the therapist. These changes include controlling the type of exposure, level of exposure, and transition to nature or stroop task environments. It also controls general settings such as audio volumes. A code snippet for changing exposure type and level can be viewed in Snippet J.1.

6.5 Software Testing

The software was tested using the same two sets of hardware used to test the eVRST application. One system included an HTC Vive HMD, GeForce GTX Titan X GPU, and Intel i7-5820k CPU. The other included an Oculus Rift S HMD, RTX 3070 GPU, and Ryzen 5 3600X CPU. Similar to the previous application, the frame rate was stable at 100 frames per second on both systems used for testing. The controls were responsive and worked as intended in both cases.

6.6 Summary

This chapter details the solution for integrating real-time signal processing and machine learning into a Unity 3D VR environment. A CNN model and information from GSR, EEG, and PPG to provide information to the therapist on the current state of anxiety of the user, which will enhance decision-making and tailor the therapy to the patient. It is predicted that real-time processing will be vital in timely intervention during therapy, positively impacting the outcome of sessions.

The first big challenge is the noise removal from signals and the calculation of brain waves, heart rate and skin conductivity in real time. The system uses FIR to apply the bandpass to the PPG and GSR signals before the calculations. The multimodality of the system using different types of measurements also helps supplement the calculations when one source of information is noisy and has to be removed.

The second challenge is using machine learning in conjunction with Unity. As anyone who has dealt with machine learning deployment before knows, it can be very messy to deal with many libraries, their dependencies, and version mismatches. Docker is used to containerise and deploy to ensure the current method will work on every machine as long as the correct image is used.

The third challenge is keeping the framerate steady while all these calculations happen. There are a few considerations that affect the framerate. The most important ones are resolution and fidelity in the VR environment and the complexity of the calculations in real-time. In the tests, it was chosen to run the servers on the same device as the Unity application for easier deployment. Representative but not photorealistic Unity scenes were utilised to ensure the

stability of the frame rate. A separate server is used to run Python calculations for heart rate and signal processing, attaining a more stable frame rate.

The current work showcases best practices for future applications of wearable sensors and machine learning in conjunction with Unity. This contribution will help make this integration smoother for future applications.

In developing this framework, a system where the therapist is in control of the patient's exposure to the stressful environment and is supported with useful information on the state of the anxiety of the user is demonstrated. This framework includes a CNN prediction of anxiety between no anxiety, mild anxiety, and severe anxiety updated every three seconds. Additionally, the therapist is presented with raw EEG, PPG, HR, and GSR information. However, the therapist has the option to disable this feature.

The accessibility of the system is one of the main goals of this thesis, making the system intuitive and easy to learn. Once the system is set up, the UI can be navigated to make transitions between different exposure scenarios without the need to change applications or set up the sensors a second time. The transition is very smooth and pain-free.

Another major part of the current work is the transferability of the decision support interface between different exposure therapy applications. Unity frameworks are utilised to ensure a smooth transition between different applications while using the same familiar interface. This makes the application transferable and scalable.

While the focus of this thesis has been on the therapeutic application of such systems, it is possible that they could be applied and used with a range of applications such as games and meditation. The non-therapeutical application of the system can be further investigated to enhance usability.

The current thesis thus far focusses on the development of a transferable system that is capable of making predictions in real time using wearable sensors and machine learning. While the usability and the stability of the system was studied, its therapeutical benefits remain unresearched. Longitudinal clinical studies are vital to confirm the theorised positive impact of such systems on therapy outcomes and retention rates.

A summary of the main findings, together with the discussion and implications of such findings, is provided in the next chapter.

Chapter 7

Conclusions and Future Directions

This chapter details, summarises, and concludes the work completed as part of this doctoral degree and discusses possible directions for future research.

7.1 Summary

The primary objective of this doctoral research was to develop a system capable of real-time classification of anxiety to enhance VR therapy with better personalisation, possibly impacting retention rates and therapy outcomes positively. The major findings of the thesis can be summarised in three parts: the establishment of ground truth for anxiety prediction, the development of real-time classification algorithms, the integration of such algorithms into a transferable VRET framework, and finally, a discussion of practical areas of use for this research.

Real-time classification of anxiety would be pointless and futile without the primary step of accurately establishing a ground truth for anxiety. This task has been an ongoing challenge in the current literature, with a distinct lack of standardisation across studies. This lack of standardisation also makes comparing the results from different studies complex.

To overcome this problem, a novel solution combining well-researched cognitive tools with a VR environment was developed. The eVRST environment, which was derived from the emotional Stroop task, was utilised for distinguishing the effects of psychological anxiety from those of task load, cognitive load, and

physical stress. It is crucial for the predictions to match the use case of exposure therapy.

Multiple environment validation methods were employed to derive meaning from the labels of no anxiety, low anxiety and mild anxiety. The results of extensive statistical and machine learning analysis showed that the emotional conditions represented separate states of anxiety based on self-report and physiological measurements.

Traditional and novel classification algorithms were meticulously applied and compared to implement the best possible solution for the dataset and use case. The CNN model outperformed all other solutions and demonstrated high accuracy and reliability in detecting anxiety states in real-time, validated by the established ground truth levels through the eStroop environment. To obtain a reliable measure of accuracy, leave-one-subject-out cross-validation was employed. Due to physiological differences between participants, there was some variety in accuracies within the sample. This nonhomogeneity confirmed the importance of multi-modal data systems combining physiological and behavioural measures.

Many past studies researching the classification of anxiety in VR employ time windows of thirty seconds or longer. Extended time windows make it challenging to derive precise information about the patient's current anxiety state. Therefore, in the current thesis, 5-second time windows were used to ensure precise and reliable information. A minimal time window is predicted to enhance the efficacy of tailored VR environments. The high reliability of the system underlines the potential for seamless objective feedback in therapeutic settings and more.

Past research in VRET applications shows a lack of unity in the user interface, user experience, and exposure settings. There are heterogeneous applications per exposure type, and no unified system can be applied through different therapy systems while also providing objective feedback. Current research makes a significant contribution to the field by developing a transferable, widely deployable and scalable framework for deploying virtual reality exposure therapy environments.

The current system comprises two major components: the exposure environment and the therapist's control panel. With the help of seamless physiological feedback and visualised data, the therapist can manipulate and

tailor the exposure environments. While this study tested only two environments, the framework system allows for the addition of unlimited exposure scenarios. These advancements will increase the engagement and enjoyment of the treatment, aiming to positively impact its effectiveness.

7.2 Application areas

The practical applications of this work extend beyond exposure therapy settings. The real-time anxiety prediction system can potentially be used in education, training, and entertainment. Objective feedback in VR is vital for any field that seeks enhancement of the experience, which applies to many fields. For instance, tailoring applications such as education and training will improve engagement and increase the benefit received from applications. Similarly, the gaming industry can improve the challenge and enjoyment aspects of games by utilising such information. The flexibility of the machine learning algorithm allows it to be deployed in extensive VR environments, making an impact on health care, education, entertainment, and more.

7.3 Virtual Reality Mental Health Interventions Limitations

Personalised VR therapy shows promise to extend the reach of mental health interventions with personalised, accessible solutions [EM21]. However, there are some shortcomings and challenges that arise from VR mental health care. Firstly, the technology used to deliver this type of therapy can be challenging to obtain in some areas[AP14]. Nowadays, affordable head-mounted displays exist for use in health care. However, the costs can add up when including specialised biosensors and upkeep of such equipment as new versions become available. These devices can also pose a challenge for patients who are not familiar with the technology [ALBY19]. Some patients can struggle using HMDs alongside their vision-correcting glasses, creating a challenge in viewing the display. Similarly, this type of therapy can be complex for patients with severe motion sickness unless designed accordingly.

Additionally, some limitations can arise with the large-scale adoption of this technology. For instance, mental health professionals would need to be trained in the use of such technology, which can be time-consuming and costly. Ethical

concerns can surface from the use of sensitive information within technology without the existence of regulatory bodies regarding mental health technology.

While there are many research studies investigating the efficacy of VR therapy, the scale and sample size of these studies are limited [EM21]. There are no global studies that study the effectiveness of these interventions, which might make the results biased towards some cultures. There also is a limited number of longitudinal studies. Most research focusses on the short-term effectiveness of the therapy method, and as the technology is relatively new, there are no discussions of long-term effects yet.

7.4 Recommendations

The research undertaken in this thesis shows the potential for significant improvements to mental health care and more. This section explores and identifies promising areas for future research. It also summarises key recommendations for future research to be undertaken in the area to increase the quality of research.

The eVRST application supplies three conditions that were identifiably distinct from each other. However, no experiments were conducted on whether the range covered in the ground truth represents the full range of emotion. The conditions developed in this thesis must be compared with their intended use case to investigate their correspondence. Additionally, benchmarking with other standard methods, such as IAPS and VR content, would aid in understanding where the conditions stand compared to previous literature.

Extending the dataset and running the experiment on diverse cultures and age groups, as well as the ones in the sample, is also essential. While this study focusses on the brevity of the experience for integration into the final product, longer experiments where each condition is experienced in a separate session on different days will improve the accuracy of the results.

The research undertaken during this PhD showed that it is crucial to use leave-one-subject-out cross-validation to identify limitations. While K-fold validation is faster and has similar end results, it fails to report individual differences. It should further be studied why some participants do not physiologically respond to stimuli the same way and how the algorithms can be enhanced for these participants.

This thesis uses easily accessible, commercial methods for collecting and processing physiological data. As a result, there were no advancements in more ergonomic or noise-compensated methods for collecting heart rate, skin conductivity and EEG. In future studies, developing systems that will allow for unrestricted movement within the VR while compensating for movement artefacts is essential. Additionally, techniques for more accurately processing dry EEG nodes could enhance both accuracy and therapy retention rate.

HRV is a crucial metric for arousal detection, which has been omitted due to the time it takes to calculate accurately from PPG recordings. Advancements in this area would greatly benefit any real-time research for emotion detection.

There is also a set of recommendations that need to be followed for the development of efficient VR-based mental health solutions. First and foremost, patients and clinicians need to be on the frontline of the development of such solutions to ensure they meet the requirements and needs of patients. Ethical guidelines must be followed when sensitive health information is being used in the applications. Any information must be recorded with the patient's informed consent. This information must also be encrypted and anonymised. Due diligence must be given to the hardware requirements of the applications, ensuring the low cost of hardware required to make it accessible to a broader population.

The VRET framework is a minimal working prototype at its current stage. Improvements are necessary to make it into a viable product. More exposure environments need to be developed per the proposed Unity framework. Such exposure environments would need to be developed in a team of computer scientists and psychologists working together.

Currently, starting the exposure is up to the therapist. It would be interesting to investigate methods of automatically detecting initial levels of exposure suited to the user. While The UI is complete, improvements can be made to make it more user-friendly. For such, user research must be conducted in collaboration with the target audience, therapists and patients.

Most importantly, large-scale, double-blind, longitudinal studies need to be conducted to research the benefits of using real-time classification compared to the traditional VRET environments. It is hypothesised that the increased personalisation will positively impact therapy outcomes.

7.5 Final Words

In conclusion, this doctoral research makes significant contributions to the field of HCI by developing a generalisable real-time anxiety classification method and integrating this into a transferable framework of VRET. The results highlight the potential of a continuous feedback framework not only for exposure therapy but also for numerous applications and fields.

References

[ACL04] Daniel Algom, Eran Chajut, and Shlomo Lev. A rational look at the emotional stroop phenomenon: a generic slowdown, not a stroop effect. *Journal of experimental psychology: General*, 133(3):323, 2004.

[ADT21] Arushi, Roberto Dillon, and Ai Ni Teoh. Real-time stress detection model and voice analysis: An integrated vr-based game for training public speaking skills. In *2021 IEEE Conference on Games (CoG)*, pages 1–4, Aug 2021.

[AEM22] Lou Ancillon, Mohamed Elgendi, and Carlo Menon. Machine learning for anxiety detection using biosignals: A review. *Diagnostics*, 12(8), 2022.

[AI21] Murf AI. Murf voice over studio. `https://murf.ai/`, 2021. Accessed: 2022-03-13.

[AJ17] Ayten Ozge Akmandor and Niraj K Jha. Keep the stress away with soda: Stress detection and alleviation system. *IEEE Transactions on Multi-Scale Computing Systems*, 3(4):269–282, 2017.

[ALBY19] Juleimar Soares Coelho de Amorim, Renata Cristine Leite, Renata Brizola, and Cristhiane Yumi Yonamine. Virtual reality therapy for rehabilitation of balance in the elderly: a systematic review and meta-analysis. *Advances in rheumatology*, 58:18, 2019.

[ALEL+18] Jordi Alonso, Zhaorui Liu, Sara Evans-Lacko, Ekaterina Sadikova, Nancy Sampson, Somnath Chatterji, Jibril Abdulmalik, Sergio Aguilar-Gaxiola, Ali Al-Hamzawi, Laura H Andrade, et al. Treatment gap for anxiety disorders is global:

Results of the world mental health surveys in 21 countries. *Depression and anxiety*, 35(3):195–208, 2018.

[Ame18a] American Psychological Association. Generalized anxiety disorder, apa dictionary of psychology. `https://dictionary.apa.org/generalized-anxiety-disorder`, 2018. Accessed: 2024-02-02.

[Ame18b] American Psychological Association. Social phobia, apa dictionary of psychology. `https://dictionary.apa.org/social-phobia`, 2018. Accessed: 2024-02-02.

[Ame19] American Psychological Association. What's the difference between stress and anxiety? `https://www.apa.org/topics/stress/anxiety-difference`, 2019. Accessed: 2022-08-20.

[Ame22] American Psychological Association. Anxiety. `https://www.apa.org/topics/anxiety`, 2022. Accessed: 2023-04-20.

[Ame23a] American Psychological Association. Panic disorder, apa dictionary of psychology. `https://dictionary.apa.org/panic-disorder`, 2023. Accessed: 2024-02-02.

[Ame23b] American Psychological Association. Specific phobia, apa dictionary of psychology. `https://dictionary.apa.org/specific-phobia`, 2023. Accessed: 2024-02-02.

[AOW+13] Kirstin Aschbacher, Aoife O'Donovan, Owen M Wolkowitz, Firdaus S Dhabhar, Yali Su, and Elissa Epel. Good stress, bad stress and oxidative stress: insights from anticipatory cortisol reactivity. *Psychoneuroendocrinology*, 38(9):1698–1708, 2013.

[AP14] Ahmed Amer and Phillip Peralez. Affordable altered perspectives: Making augmented and virtual reality technology accessible. In *IEEE global humanitarian technology conference (GHTC 2014)*, pages 603–608. IEEE, 2014.

[APE+13] Page L Anderson, Matthew Price, Shannan M Edwards,

Mayowa A Obasaju, Stefan K Schmertz, Elana Zimand, and Martha R Calamaras. Virtual reality exposure therapy for social anxiety disorder: a randomized controlled trial. *Journal of consulting and clinical psychology*, 81(5):751, 2013.

[ARE+13] Christina M Armstrong, Greg M Reger, Joseph Edwards, Albert A Rizzo, Christopher G Courtney, and Thomas D Parsons. Validity of the virtual reality stroop task (vrst) in active duty military. *Journal of Clinical and Experimental Neuropsychology*, 35(2):113–123, 2013.

[ASR+03] H Harry Asada, Phillip Shaltis, Andrew Reisner, Sokwoo Rhee, and Reginald C Hutchinson. Mobile monitoring with wearable photoplethysmographic biosensors. *IEEE engineering in medicine and biology magazine*, 22(3):28–40, 2003.

[BA19] Amanda A Benbow and Page L Anderson. A meta-analytic examination of attrition in virtual reality exposure therapy for anxiety disorders. *Journal of Anxiety Disorders*, 61:18–26, 2019.

[BAM+20] Joscha Boehnlein, Luisa Altegoer, Nina Kristin Muck, Kati Roesmann, Ronny Redlich, Udo Dannlowski, and Elisabeth J Leehr. Factors influencing the success of exposure therapy for specific phobia: A systematic review. *Neuroscience & Biobehavioral Reviews*, 108:796–820, 2020.

[BB11] Raymond Bruyer and Marc Brysbaert. Combining speed and accuracy in cognitive psychology: Is the inverse efficiency score (ies) a better dependent variable than the mean reaction time (rt) and the percentage of errors (pe)? *Psychologica Belgica*, 51(1):5–13, 2011.

[BBM01] John Bound, Charles Brown, and Nancy Mathiowetz. Measurement error in survey data. In *Handbook of econometrics*, volume 5, pages 3705–3843. Elsevier, 2001.

[BBW17] Sophie Michaelis Borwin Bandelow and Dirk Wedekind. Treatment of anxiety disorders. *Dialogues in Clinical Neuroscience*, 19(2):93–107, 2017.

[BDG+10] Adrian Burns, Emer P Doheny, Barry R Greene, Timothy Foran, Daniel Leahy, Karol O'Donovan, and Michael J

McGrath. Shimmer™: an extensible platform for physiological signal capture. In *2010 annual international conference of the IEEE engineering in medicine and biology*, pages 3759–3762. IEEE, 2010.

[BDR+17] Stephane Bouchard, Stephanie Dumoulin, Genevieve Robillard, Tanya Guitard, Evelyne Klinger, Helene Forget, Claudie Loranger, and Francois Xavier Roucaut. Virtual reality compared with in vivo exposure in the treatment of social anxiety disorder: a three-arm randomised controlled trial. *The British Journal of Psychiatry*, 210(4):276–283, 2017.

[BFF21] N. Bu, M. Fukami, and O. Fukuda. Pattern recognition of mental stress levels from differential rri time series using lstm networks. *LifeTech 2021 - 2021 IEEE 3rd Global Conference on Life Sciences and Technologies*, pages 408–411, 2021.

[BHvE15] Willem-Paul Brinkman, Allart RD Hoekstra, and René van EGMOND. The effect of 3d audio and other audio techniques on virtual reality experience. *Annual Review of Cybertherapy and Telemedicine 2015*, pages 44–48, 2015.

[BHWH+16] Moshe Shay Ben-Haim, Paul Williams, Zachary Howard, Yaniv Mama, Ami Eidels, and Daniel Algom. The emotional stroop task: assessing cognitive performance under exposure to emotional content. *JoVE (Journal of Visualized Experiments)*, (112):e53720, 2016.

[BK10a] Mathias Benedek and Christian Kaernbach. A continuous measure of phasic electrodermal activity. *Journal of neuroscience methods*, 190(1):80–91, 2010.

[BK10b] Mathias Benedek and Christian Kaernbach. Decomposition of skin conductance data by means of nonnegative deconvolution. *psychophysiology*, 47(4):647–658, 2010.

[BL99] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology . . . , 1999.

[BMM+20] O. Bălan, G. Moise, A. Moldoveanu, F. Moldoveanu, and

M. Leordeanu. Classifying the levels of fear by means of machine learning techniques and vr in a holonic-based system for treating phobias. experiments and results. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12191 Lncs:357–372, 2020.

[BNK+11] Anne-Marie Brouwer, Mark A Neerincx, Victor Kallen, Leslie van der Leer, and Michiel ten Brinke. Eeg alpha asymmetry, heart rate variability and cortisol in response to virtual reality induced stress. *J. Cyberther. Rehabil*, 4(1):21–34, 2011.

[BR13] Isabelle Blanchette and Anne Richards. Is emotional stroop interference linked to affective responses? evidence from skin conductance and facial electromyography. *Emotion*, 13(1):129, 2013.

[Bre04] J Douglas Bremner. Brain imaging in anxiety disorders. *Expert Review of Neurotherapeutics*, 4(2):275–284, 2004.

[BSSM14] Vicki Bitsika, Christopher F. Sharpley, John A. Sweeney, and James R. McFarlane. Hpa and sam axis responses as correlates of self- vs parental ratings of anxiety in boys with an autistic disorder. *Physiology and Behavior*, 127:1–7, March 2014.

[BST+02] Borwin Bandelow, Corinna Späth, Gabriel Álvarez Tichauer, Andreas Broocks, Göran Hajak, and Eckart Rüther. Early traumatic life events, parental attitudes, family history, and birth risk factors in patients with panic disorder. *Comprehensive psychiatry*, 43(4):269–278, 2002.

[BSY+17] Nicolas Bourdillon, Laurent Schmitt, Sasan Yazdani, Jean-Marc Vesin, and Grégoire P Millet. Minimal window duration for accurate hrv recording in athletes. *Frontiers in neuroscience*, 11:456, 2017.

[CAN05] Mihaly Csikszentmihalyi, Sami Abuhamdeh, and Jeanne Nakamura. Flow. *Handbook of competence and motivation*, pages 598–608, 2005.

[CC67] CG Costello and Andrew L Comrey. Scales for measuring

depression and anxiety. *The Journal of psychology*, 66(2):303–313, 1967.

[CCEE19] Yekta Said Can, Niaz Chalabianloo, Deniz Ekiz, and Cem Ersoy. Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study. *Sensors*, 19(8):1849, 2019.

[CCF+18] Peter H Charlton, Patrick Celka, Bushra Farukh, Phil Chowienczyk, and Jordi Alastruey. Assessing mental stress from the photoplethysmogram: a numerical study. *Physiological measurement*, 39(5):054001, 2018.

[CCN+12] D. Carneiro, J.C. Castillo, P. Novais, A. Fernández-Caballero, and J. Neves. Multimodal behavioral analysis for non-invasive stress detection. *Expert Systems with Applications*, 39(18):13376–13389, 2012.

[CHO+17] D. Cho, J. Ham, J. Oh, J. Park, S. Kim, N.-K. Lee, and B. Lee. Detection of stress levels from biosignals measured in virtual reality environments using a kernel-based extreme learning machine. *Sensors (Switzerland)*, 17(10), 2017.

[CKY20] Eunhee Chang, Hyun Taek Kim, and Byounghyun Yoo. Virtual reality sickness: a review of causes and measurements. *International Journal of Human–Computer Interaction*, 36(17):1658–1682, 2020.

[Cri19] Critical Appraisal Skills Programme. Casp qualitative studies checklist. `https://casp-uk.net/casp-tools-checklists/`, 2019. Accessed: 2021-05-11.

[CSLC+19a] Emily Carl, Aliza T Stein, Andrew Levihn-Coon, Jamie R Pogue, Barbara Rothbaum, Paul Emmelkamp, Gordon JG Asmundson, Per Carlbring, and Mark B Powers. Virtual reality exposure therapy for anxiety and related disorders: A meta-analysis of randomized controlled trials. *Journal of anxiety disorders*, 61:27–36, 2019.

[CSLC+19b] Emily Carl, Aliza T. Stein, Andrew Levihn-Coon, Jamie R. Pogue, Barbara Rothbaum, Paul Emmelkamp, Gordon J.G. Asmundson, Per Carlbring, and Mark B. Powers. Virtual

reality exposure therapy for anxiety and related disorders: A meta-analysis of randomized controlled trials. *Journal of Anxiety Disorders*, 61:27–36, 2019. Virtual reality applications for the anxiety disorders.

[CW10] Carlos M. Coelho and Guy Wallis. Deconstructing acrophobia: physiological and psychological precursors to developing a fear of heights. *Depression and Anxiety*, 27(9):864–870, 2010.

[CWHW09] Carlos M. Coelho, Allison M. Waters, Trevor J. Hine, and Guy Wallis. The use of virtual reality in acrophobia research and treatment. *Journal of Anxiety Disorders*, 23(5):563–574, 2009.

[CXC19] Mingliang Cao, Tianhua Xie, and Zebin Chen. Wearable sensors and equipment in vr games: a review. *Transactions on Edutainment XV*, pages 3–12, 2019.

[CXM+16] Rossana Castaldo, William Xu, Paolo Melillo, Leandro Pecchia, Lorena Santamaria, and C James. Detection of mental stress due to oral academic examination via ultra-short-term hrv analysis. In *2016 38th Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 3805–3808. IEEE, 2016.

[CZY+17] Lan-lan Chen, Yu Zhao, Peng-fei Ye, Jian Zhang, and Jun-zhong Zou. Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers. *Expert Systems with Applications*, 85:279–291, 2017.

[DAP+15] Julia Diemer, Georg W Alpers, Henrik M Peperkorn, Youssef Shiban, and Andreas Mühlberger. The impact of perception and presence on emotional reactions: a review of research in virtual reality. *Frontiers in psychology*, 6:26, 2015.

[Dav68] Gerald C. Davison. Systematic desensitization as a counterconditioning process. *Journal of Abnormal Psychology*, 73(2):91–99, 1968.

[DBH+15] Tara Donker, Matthijs Blankers, Erik Hedman, Brjann Ljotsson, Katherine Petrie, and Helen Christensen. Economic evaluations of internet interventions for mental health: a

systematic review. *Psychological medicine*, 45(16):3357–3376, 2015.

[DC07] Andrew Dekker and Erik Champion. Please biofeed the zombies: enhancing the gameplay and display of a horror game using biofeedback. In *DiGRA'07-Proceedings of the 2007 DiGRA International Conference: Situated Play*, pages 550–558, 2007.

[DDP09] Katarina Dedovic, Catherine D'Aguiar, and Jens C Pruessner. What stress does to your brain: a review of neuroimaging studies. *The Canadian Journal of Psychiatry*, 54(1):6–15, 2009.

[dSSÁCDP11] Alberto de Santos Sierra, Carmen Sánchez Ávila, Javier Guerra Casanova, and Gonzalo Bailador Del Pozo. A stress-detection system based on physiological signals and fuzzy logic. *IEEE transactions on industrial electronics*, 58(10):4857–4865, 2011.

[EBM18] William W Eaton, O Joseph Bienvenu, and Beyon Miloyan. Specific phobias. *The Lancet Psychiatry*, 5(8):678–686, 2018.

[EH11] Tiina M Eilola and Jelena Havelka. Behavioural and physiological responses to the emotional and taboo stroop tasks in native and non-native speakers of english. *International Journal of Bilingualism*, 15(3):353–369, 2011.

[EM21] Paul MG Emmelkamp and Katharina Meyerbröker. Virtual reality therapy in mental health. *Annual review of clinical psychology*, 17:495–519, 2021.

[Eng22] Shimmer Engineering. Shimmer c api wiki. `https://github.com/ShimmerEngineering/Shimmer-C-API/wiki/Verisense-General`, 2022. Accessed: 2022-12-06.

[FAB+04] Carlo Faravelli, Luca Abrardi, Daniela Bartolozzi, Cristiana Cecchi, Fiammetta Cosci, Donato D'Adamo, Beatrice Lo Iacono, Claudia Ravaldi, Maria Alessandra Scarpato, Elisabetta Truglia, et al. The sesto fiorentino study: Background, methods and preliminary results: Lifetime prevalence of psychiatric disorders in an italian community sample using clinical interviewers. *Psychotherapy and Psychosomatics*, 73(4):216–225, 2004.

[GAT+20] Gerald Gartlehner, Lisa Affengruber, Viktoria Titscher, Anna Noel-Storr, Gordon Dooley, Nicolas Ballarini, and Franz König. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. *Journal of clinical epidemiology*, 121:20–28, may 2020.

[Gav20] Paul Gavrikov. visualkeras. `https://github.com/paulgavrikov/visualkeras`, 2020.

[GC89] Ian H Gotlib and Douglas B Cane. Self-report assessment of depression and anxiety. *Anxiety and depression: Distinctive and overlapping features*, pages 131–169, 1989.

[GCS+14] Andrea Gaggioli, Pietro Cipresso, Silvia Serino, Giovanni Pioggia, Gennaro Tartarisco, Giovanni Baldus, Daniele Corda, Marcello Ferro, Nicola Carbonaro, Alessandro Tognetti, et al. A decision support system for real-time stress detection during virtual reality exposure. In *MMVR*, volume 2014, pages 114–120, 2014.

[GGG+19] Giorgos Giannakakis, Dimitris Grigoriadis, Katerina Giannakaki, Olympia Simantiraki, Alexandros Roniotis, and Manolis Tsiknakis. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, 2019.

[GIM22] GIMP. Gimp - gnu image manipulation program. `https://www.gimp.org/`, 2022. Accessed: 2022-12-20.

[GLC+13] V.V. Galatenko, E.D. Livshitz, A.M. Chernorizov, Y.P. Zinchenko, A.V. Galatenko, V.M. Staroverov, S.A. Isaychev, V.V. Lebedev, G.Y. Menshikova, A.N. Gusev, E.M. Lobacheva, R.F. Gabidullina, V.E. Podol'skii, and V.A. Sadovnichy. Automated real-time classification of functional states: Significance of individual tuning stage. *Psychology in Russia: State of the Art*, 6(3):41–48, 2013.

[GMSSMFL22] Manuel Gil-Martin, Ruben San-Segundo, Ana Mateos, and Javier Ferreiros-Lopez. Human stress detection with wearable sensors using convolutional neural networks. *IEEE Aerospace and Electronic Systems Magazine*, 37(1):60–70, 2022.

[GP21] Shruti Gedam and Sanchita Paul. A review on mental stress detection using wearable sensors and machine learning techniques. *IEEE Access*, 9:84045–84066, 2021.

[GPHKS+01] Azucena Garcia-Palacios, Hunter G Hoffman, Sheree Kwong See, AMY Tsai, and Cristina Botella. Redefining therapeutic success with virtual reality exposure therapy. *CyberPsychology & behavior*, 4(3):341–348, 2001.

[GPT+12] Andrea Gaggioli, Giovanni Pioggia, Gennaro Tartarisco, Giovanni Baldus, Marcello Ferro, Pietro Cipresso, Silvia Serino, Andrei Popleteev, Silvia Gabrielli, Rosa Maimone, et al. A system for automatic detection of momentary stress in naturalistic settings. *Annual Review of Cybertherapy and Telemedicine*, pages 182–186, 2012.

[GRDP+18] Lívea Dornela Godoy, Matheus Teixeira Rossignoli, Polianna Delfino-Pereira, Norberto Garcia-Cairasco, and Eduardo Henrique de Lima Umeoka. A comprehensive overview on stress neurobiology: Basic concepts and clinical implications, July 2018.

[GWM+19] Stefan Gradl, Markus Wirth, Nico Mächtlinger, Romina Poguntke, Andrea Wonner, Nicolas Rohleder, and Bjoern M Eskofier. The stroop room: A virtual reality-enhanced stroop test. In *25th ACM Symposium on Virtual Reality Software and Technology*, pages 1–12, 2019.

[HCOL17] J. Ham, D. Cho, J. Oh, and B. Lee. Discrimination of multiple stress levels in virtual reality environments using heart rate variability. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 3989–3992, 2017.

[HEK20] C. Hirt, M. Eckard, and A. Kunz. Stress generation and non-intrusive measurement in virtual environments using eye tracking. *Journal of Ambient Intelligence and Humanized Computing*, 11(12):5977–5989, 2020.

[HHK18] Nathalie Hauk, Joachim Hüffmeier, and Stefan Krumm. Ready to be a silver surfer? a meta-analysis on the relationship

between chronological age and technology acceptance. *Computers in Human Behavior*, 84:304–319, 2018.

[HLL+19] Jiayuan He, Ke Li, Xiaoli Liao, Ping Zhang, and Ning Jiang. Real-time detection of acute cognitive stress using a convolutional neural network from electrocardiographic signal. *IEEE Access*, 7:42710–42717, 2019.

[HMP+13] W. Handouzi, C. Maaoui, A. Pruski, A. Moussaoui, and Y. Bendiouis. Short-term anxiety recognition induced by virtual reality exposure for phobic people. *Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013*, pages 3145–3150, 2013.

[HMPM14a] W. Handouzi, C. Maaoui, A. Pruski, and A. Moussaoui. Objective model assessment for short-term anxiety recognition from blood volume pulse signal. *Biomedical Signal Processing and Control*, 14(1):217–227, 2014.

[HMPM14b] W. Handouzi, C. Maaoui, A. Pruski, and A. Moussaoui. Short-term anxiety recognition from blood volume pulse signal. *2014 IEEE 11th International Multi-Conference on Systems, Signals and Devices, SSD 2014*, 2014.

[HPM+05] John M Hettema, Carol A Prescott, John M Myers, Michael C Neale, and Kenneth S Kendler. The structure of genetic and environmental risk factors for anxiety disorders in men and women. *Archives of general psychiatry*, 62(2):182–189, 2005.

[HRK+94] Larry F. Hodges, Barbara O. Rothbaum, Rob Kooper, Dan Opdyke, Thomas C. Meyer, and Johannes J. de Graaff. Presence as the defining factor in a vr application. *Computer Science*, 1994.

[HRW19] Jamiah Hyde, Katherine M Ryan, and Allison M Waters. Psychophysiological markers of fear and anxiety. *Current Psychiatry Reports*, 21:1–10, 2019.

[HSVE+19] Hanne Huygelier, Brenda Schraepen, Raymond Van Ee, Vero Vanden Abeele, and Céline R Gillebert. Acceptance of immersive head-mounted virtual reality in older adults. *Scientific reports*, 9(1):4519, 2019.

[IKJM+19] Young Im Kim, Seo-Yeon Jung, Seulki Min, Eunbi Seol, Sungho Seo, Ji-Won Hur, Dooyoung Jung, Heon-Jeong Lee, Sungkil Lee, Gerard J Kim, et al. Visuo-haptic-based multimodal feedback virtual reality solution to improve anxiety symptoms: a proof-of-concept study. *Psychiatry Investigation*, 16(2):167, 2019.

[Inc24] Empatica Inc. E4 wristband - real-time physiological signals. `https://www.empatica.com/research/e4/`, 2024. Accessed: 2024-01-06.

[Ins20] AR Insider. New report: 19% of u.s. adults have tried vr. `https://arinsider.co/2020/05/20/new-report-19-of-u-s-adults-have-tried-vr/`, 2020. Accessed: 2022-20-12.

[IRN+20] S. Ishaque, A. Rueda, B. Nguyen, N. Khan, and S. Krishnan. Physiological signal analysis and classification of stress from virtual reality video game. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2020-July:867–870, 2020.

[JKJ20] S.Y. Ji, S.Y. Kang, and H.J. Jun. Deep-learning-based stress-ratio prediction model using virtual reality with electroencephalography data. *Sustainability (Switzerland)*, 12(17), 2020.

[JRJ66] Arthur R Jensen and William D Rohwer Jr. The stroop color-word test: a review. *Acta psychologica*, 25:36–93, 1966.

[KAK21] J. Kritikos, G. Alevizopoulos, and D. Koutsouris. Personalized virtual reality human-computer interaction for psychiatric and neurological illnesses: A dynamically adaptive virtual reality environment that changes according to real-time feedback from electrophysiological signal responses. *Frontiers in Human Neuroscience*, 15, 2021.

[KBB96] Merel Kindt, Dick Bierman, and Jos F Brosschot. Stroop versus stroop: Comparison of a card format and a single-trial format of the standard color-word stroop task and the emotional stroop task. *Personality and individual Differences*, 21(5):653–661, 1996.

[KBD+05] Ronald C Kessler, Patricia Berglund, Olga Demler, Robert Jin, Kathleen R Merikangas, and Ellen E Walters. Lifetime prevalence and age-of-onset distributions of dsm-iv disorders in the national comorbidity survey replication. *Archives of general psychiatry*, 62(6):593–602, 2005.

[KEB+04] Merel Krijn, Paul M.g Emmelkamp, Roeline Biemond, Claudius De Wilde De Ligny, Martijn J Schuemie, and Charles A.p.g Van Der Mast. Treatment of acrophobia in virtual reality: The role of immersion and presence. *Behaviour Research and Therapy*, 42(2):229–239, 2004.

[KEH+16] Isabel L Kampmann, Paul MG Emmelkamp, Dwi Hartanto, Willem-Paul Brinkman, Bonne JH Zijlstra, and Nexhmedin Morina. Exposure to virtual social interactions in the treatment of social anxiety disorder: A randomized controlled trial. *Behaviour research and therapy*, 77:147–156, 2016.

[KEÓ+07] Merel Krijn, Paul MG Emmelkamp, Ragnar P Ólafsson, Manon Bouwman, Lucas J Van Gerwen, Philip Spinhoven, Martijn J Schuemie, and Charles APG Van der Mast. Fear of flying treatment methods: virtual reality exposure vs. cognitive behavioral therapy. *Aviation, space, and environmental medicine*, 78(2):121–128, 2007.

[KEOB04] Merel Krijn, Paul MG Emmelkamp, Ragnar P Olafsson, and Roeline Biemond. Virtual reality exposure therapy of anxiety disorders: A review. *Clinical psychology review*, 24(3):259–281, 2004.

[KGS12] Konstantina Kilteni, Raphaela Groten, and Mel Slater. The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments*, 21(4):373–387, 2012.

[KHBK20] H. Kusano, Y. Horiguchi, Y. Baba, and H. Kashima. Stress prediction from head motion. *Proceedings - 2020 IEEE 7th International Conference on Data Science and Advanced Analytics, DSAA 2020*, pages 488–495, 2020.

[KHL+20] S. Klingner, Z. Han, Y. Liu, F. Fan, B. Altakrouri, B. Michel, J. Weiss, A. Sridhar, and S.M. Chau. Firefighter virtual reality

simulation for personalized stress detection. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12325 Lnai:343–347, 2020.

[KMY14] P Karthikeyan, M Murugappan, and S Yaacob. Analysis of stroop color word test-based human stress detection using electrocardiography and heart rate variability signals. *Arabian Journal for Science and Engineering*, 39(3):1835–1847, 2014.

[KO20] Kelly A Knowles and Bunmi O Olatunji. Specificity of trait anxiety in anxiety and depression: Meta-analysis of the state-trait anxiety inventory. *Clinical psychology review*, 82:101928, 2020.

[KSGW15] Anna L. Kratz, Stephen G. Schilling, Jenna Goesling, and David A. Williams. Development and initial validation of a brief self-report measure of cognitive dysfunction in fibromyalgia. *The Journal of Pain*, 16(6):527–536, 2015.

[KSZ21] D. Kamińska, K. Smółka, and G. Zwoliński. Detection of mental stress through eeg signal in virtual reality environment. *Electronics (Switzerland)*, 10(22), Nov 2021.

[KTC01] Einar Kringlen, Svenn Torgersen, and Victoria Cramer. A norwegian psychiatric epidemiological study. *American journal of psychiatry*, 158(7):1091–1098, 2001.

[LD18] Yun Liu and Siqing Du. Psychological stress level detection based on electrodermal activity. *Behavioural brain research*, 341:50–53, 2018.

[Lev15] Eugene E Levitt. *The psychology of anxiety*. Routledge, 2015.

[LIF+09] Marc Lambooij, Wijnand IJsselsteijn, Marten Fortuin, Ingrid Heynderickx, et al. Visual discomfort and visual fatigue of stereoscopic displays: a review. *Journal of imaging science and technology*, 53(3):30201–1, 2009.

[LKH+21] Lenka Lhotska, Patrik Kutilek, Jan Husak, Anna Francova, Jan Hejda, Jakub Stejskal, Martin Kotek, Jan Hybl, Pavla Stopkova, and Iveta Fajnerova. Virtual reality for exposure

therapy. In *2021 14th International Conference on Human System Interaction (HSI)*, pages 1–6, 2021.

[LSL+17] Y. Liu, O. Sourina, H.P. Liew, H.S. Salem Chandrasekaran, D. Konovessis, G. Krishnan, and H.E. Ang. Human factors evaluation in maritime virtual simulators using mobile eeg-based neuroimaging. *Advances in Transdisciplinary Engineering*, 5:261–268, 2017.

[LSZM21] Leonidas Liakopoulos, Nikolaos Stagakis, Evangelia I Zacharaki, and Konstantinos Moustakas. Cnn-based stress and emotion recognition in ambulatory settings. In *2021 12th international conference on information, intelligence, systems & applications (IISA)*, pages 1–8. IEEE, 2021.

[LYS+17] Lan Li, Fei Yu, Dongquan Shi, Jianping Shi, Zongjun Tian, Jiquan Yang, Xingsong Wang, and Qing Jiang. Application of virtual reality technology in clinical medicine. *American journal of translational research*, 9(9):3867, 2017.

[LZZJ05] Wenhui Liao, Weihong Zhang, Zhiwei Zhu, and Qiang Ji. A real-time human stress monitoring system using dynamic bayesian network. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)-workshops*, pages 70–70. IEEE, 2005.

[Mac91] Colin M MacLeod. Half a century of research on the stroop effect: an integrative review. *Psychological bulletin*, 109(2):163, 1991.

[MB92] Theresa M. Marteau and Hilary Bekker. The development of a six-item short-form of the state scale of the spielberger state–trait anxiety inventory (stai). *British Journal of Clinical Psychology*, 31(3):301–306, 1992.

[McC07] Laurie Kelly McCorry. Physiology of the autonomic nervous system. *American journal of pharmaceutical education*, 71(4), 2007.

[McE07] Bruce S McEwen. Physiology and neurobiology of stress and adaptation: central role of the brain. *Physiological reviews*, 87(3):873–904, 2007.

[MG16] V Mitrousia and O Giotakos. Virtual reality therapy in anxiety disorders. *Psychiatrike= Psychiatriki*, 27(4):276–286, 2016.

[MH13] Amanda S Morrison and Richard G Heimberg. Social anxiety and social anxiety disorder. *Annual review of clinical psychology*, 9(1):249–274, 2013.

[MH20] Luke A. McGuinness and Julian P. T. Higgins. Risk-of-bias visualization (robvis): An r package and shiny web app for visualizing risk-of-bias assessments. *Research Synthesis Methods*, n/a(n/a), 2020.

[MLS+05] Gin S Malhi, Jim Lagopoulos, Perminder S Sachdev, Belinda Ivanovski, and Ron Shnier. An emotional stroop functional mri study of euthymic bipolar disorder. *Bipolar disorders*, 7:58–69, 2005.

[MMKE13] K Meyerbroeker, N Morina, GA Kerkhof, and PMG Emmelkamp. Virtual reality exposure therapy does not provide any additional value in agoraphobic patients: a randomized controlled trial. *Psychotherapy and psychosomatics*, 82(3):170–176, 2013.

[MMT21] Deniz Mevlevioğlu, David Murphy, and Sabin Tabirca. Real-time anxiety prediction in virtual reality exposure therapy. In *ACM International Conference on Interactive Media Experiences*, IMX '21, New York, NY, USA, 2021. Association for Computing Machinery.

[MPL+21] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, February 2021.

[MTM21a] Deniz Mevlevioğlu, Sabin Tabirca, and David Murphy. Emotional virtual reality stroop task: Pilot design. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, pages 1–3, 2021.

[MTM21b] Deniz Mevlevioğlu, , Sabin Tabirca, and David Murphy. Visual respiratory feedback in virtual reality exposure therapy: A

pilot study. In *ACM International Conference on Interactive Media Experiences*, IMX '21, page 1–6, New York, NY, USA, 2021. Association for Computing Machinery.

[MTM22a] Deniz Mevlevioğlu, Sabin Tabirca, and David Murphy. Emotional virtual reality stroop task: an immersive cognitive test. In *Proceedings of the 2022 ACM International Conference on Interactive Media Experiences*, pages 387–390, 2022.

[MTM22b] Deniz Mevlevioğlu, Sabin Tabirca, and David Murphy. Real-time anxiety prediction in virtual reality therapy: research proposal. In *Proceedings of the 13th ACM Multimedia Systems Conference*, MMSys '22, page 352–356, New York, NY, USA, 2022. Association for Computing Machinery.

[MTM23a] Deniz Mevlevioğlu, Sabin Tabirca, and David Murphy. Anxiety classification in virtual reality using biosensors: A mini scoping review. *Plos one*, 18(7):e0287984, 2023.

[MTM23b] Deniz Mevlevioğlu, Sabin Tabirca, and David Murphy. Emotional virtual reality stroop task. In *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 172–177. IEEE, 2023.

[MTM24] Deniz Mevlevioğlu, Sabin Tabirca, and David Murphy. Real-time classification of anxiety in virtual reality therapy using biosensors and a convolutional neural network. *Biosensors*, 14(3):131, 2024.

[Mye06] K. M. Myers. Different mechanisms of fear extinction dependent on length of time since fear acquisition. *Learning & Memory*, 13(2):216–223, Jan 2006.

[Neu22a] Neurosky. Myndband eeg brainwave headset. `https://store.myndplay.com/products.php?prod=48`, 2022. Accessed: 2022-12-06.

[Neu22b] Neurosky. neurosky 101 [neurosky developer docs]. `https://developer.neurosky.com/docs/doku.php?id=neurosky%5F101`, 2022. Accessed: 2022-12-06.

[NP02] Sarah Nichols and Harshada Patel. Health and safety

implications of virtual reality: a review of empirical evidence. *Applied Ergonomics*, 33(3):251–271, 2002. Fundamental Reviews in Applied Ergonomics 2002.

[OPGP+11] David Opris, Sebastian Pintea, Azucena García-Palacios, Cristina Botella, Stefan Szamosközi, and Daniel David. Virtual reality exposure therapy in anxiety disorders: a quantitative meta-analysis. *Depression and Anxiety*, 29(2):85–93, Jul 2011.

[PB18] Thomas D. Parsons and Michael D. Barnett. Virtual apartment stroop task: Comparison with computerized and traditional stroop tasks. *Journal of Neuroscience Methods*, 309:35–40, 2018.

[PCAD11] Thomas D Parsons, Christopher G Courtney, Brian Arizmendi, and Michael Dawson. Virtual reality stroop task for neurocognitive assessment. In *Medicine Meets Virtual Reality 18*, pages 433–439. IOS Press, 2011.

[PCR+13] Federica Pallavicini, Pietro Cipresso, Simona Raspelli, Alessandra Grassi, Silvia Serino, Cinzia Vigna, Stefano Triberti, Marco Villamira, Andrea Gaggioli, Giuseppe Riva, and et al. Is virtual reality always an effective stressors for exposure treatments? some insights from a controlled trial. *BMC Psychiatry*, 13(1), Nov 2013.

[PMB+21] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, and David Moher. Updating guidance for reporting systematic reviews: development of the prisma 2020 statement. *Journal of Clinical Epidemiology*, 134:103–112, 2021.

[PPM+20] L. Petrescu, C. Petrescu, O. Mitruț, G. Moise, A. Moldoveanu, F. Moldoveanu, and M. Leordeanu. Integrating biosignals measurement in virtual reality environments for anxiety detection. *Sensors (Switzerland)*, 20(24):1–32, 2020.

[PSN18] C. Park, S. Shahrdar, and M. Nojoumian. Eeg-based classification of emotional state using an autonomous vehicle simulator. *Proceedings of the IEEE Sensor Array and*

Multichannel Signal Processing Workshop, 2018-July:297–300, 2018.

[PVVBLGM21] E. Perez-Valero, M.A. Vaquero-Blasco, M.A. Lopez-Gordo, and C. Morillas. Quantitative assessment of stress through eeg during a virtual reality stress-relax session. *Frontiers in Computational Neuroscience*, 15, Jul 14 2021.

[Rac64] Stanley Rachman. The treatment of anxiety and phobic reactions by systematic desensitization psychotherapy. *Experiments in Behaviour Therapy*, page 62–70, 1964.

[RB97] Patrice Renaud and Jean-Pierre Blondin. The stress of stroop performance: Physiological and emotional responses to color–word interference, task pacing, and pacing speed. *International Journal of Psychophysiology*, 27(2):87–97, 1997.

[RBCS06] Peter P Roy-Byrne, Michelle G Craske, and Murray B Stein. Panic disorder. *The Lancet*, 368(9540):1023–1032, 2006.

[RBS+22] Muhammad Arifur Rahman, David J Brown, Nicholas Shopland, Matthew C Harris, Zakia Batool Turabee, Nadja Heym, Alexander Sumich, Brad Standen, David Downes, Yangang Xing, et al. Towards machine learning driven self-guided virtual reality exposure therapy based on arousal state detection from multimodal data. In *International Conference on Brain Informatics*, pages 195–209. Springer, 2022.

[REG+19] Simon Riches, Soha Elghany, Philippa Garety, Mar Rus-Calafell, and Lucia Valmaggia. Factors affecting sense of presence in a virtual reality social environment: a qualitative study. *Cyberpsychology, Behavior, and Social Networking*, 22(4):288–292, 2019.

[RHK+95] Barbara Olasov Rothbaum, Larry F. Hodges, Rob Kooper, Dan Opdyke, James S. Williford, and Max North. Virtual reality graded exposure in the treatment of acrophobia: A case report. *Behavior Therapy*, 26(3):547–554, 1995.

[RHK97] Barbara Olasov Rothbaum, Larry Hodges, and Rob Kooper.

Virtual reality exposure therapy. *Journal of Psychotherapy Practice & Research*, 1997.

[Riv03] Guiseppe Riva. Applications of virtual environments in medicine. *Methods of information in medicine*, 42(05):524–534, 2003.

[Riv05] Giuseppe Riva. Virtual reality in psychotherapy: Review. *CyberPsychology & Behavior*, 8(3):220–230, 2005.

[RMC+07] Giuseppe Riva, Fabrizia Mantovani, Claret Samantha Capideville, Alessandra Preziosa, Francesca Morganti, Daniela Villani, Andrea Gaggioli, Cristina Botella, and Mariano Alcañiz. Affective interactions using virtual reality: the link between presence and emotions. *Cyberpsychology & behavior*, 10(1):45–56, 2007.

[RMRF00] BO Rothbaum, EA Meadows, P Resick, and DW Foy. Cognitive-behavioral therapy. In EB Foa, TM Keane, and MJ Friedman, editors, *Effective treatments for PTSD: Practice guidelines from the International Society for Traumatic Stress Studies*, pages 320–325. Guilford Press, New York, NY, US, 2000.

[RO16] Lisa Rebenitsch and Charles Owen. Review on cybersickness in applications and visual displays. *Virtual Reality*, 20:101–125, 2016.

[RS91] Gary E Riccio and Thomas A Stoffregen. An ecological theory of motion sickness and postural instability. *Ecological psychology*, 3(3):195–240, 1991.

[RS02] Holger Regenbrecht and Thomas Schubert. Real and illusory interactions enhance presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 11(4):425–434, 2002.

[RSFAE+12] A. Riera, A. Soria-Frisch, A. Albajes-Eizagirre, P. Cipresso, C. Grau, S. Dunne, and G. Ruffini. Electro-physiological data fusion for stress detection. *Annual Review of CyberTherapy and Telemedicine*, 10:228–232, 2012.

[SAE+15] Virginia Sandulescu, Sally Andrews, David Ellis, Nicola

Bellotto, and Oscar Martinez Mozos. Stress detection using wearable physiological sensors. In *Artificial Computation in Biology and Medicine: International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2015, Elche, Spain, June 1-5, 2015, Proceedings, Part I 6*, pages 526–532. Springer, 2015.

[SAJC05] Gregory P. Strauss, Daniel N. Allen, Melinda L. Jorgensen, and Stacey L. Cramer. Test-retest reliability of standard and emotional stroop tasks: An investigation of color-word and picture-word versions. *Assessment*, 12(3):330–337, 2005. PMID: 16123253.

[Sał14] Wojciech Sałabun. Processing and spectral analysis of the raw eeg signal from the mindwave. *Przeglad Elektrotechniczny*, 90(2):169–174, 2014.

[SBD+00] MJ Schuemie, M Bruynzeel, L Drost, M Brinckman, G De Haan, PMG Emmelkamp, and CAPG Van der Mast. Treatment of acrophobia in virtual reality: A pilot study. In *Conference Proceedings Euromedia*, pages 271–275, 2000.

[SC15] Riccardo Sioni and Luca Chittaro. Stress detection using physiological sensors. *Computer*, 48(10):26–33, 2015.

[SCCC14] Benjamin D Schalet, Karon F Cook, Seung W Choi, and David Cella. Establishing a common metric for self-reported anxiety: linking the masq, panas, and gad-7 to promis anxiety. *Journal of anxiety disorders*, 28(1):88–96, 2014.

[Sen22] Shimmer Sensing. Shimmer3 gsr+ unit - shimmer wearable sensor technology. `https://shimmersensing.com/product/shimmer3-gsr-unit/`, 2022. Accessed: 2022-12-06.

[SGL+83] Charles Spielberger, Richard Gorsuch, Robert Lushene, PR Vagg, and Gerard Jacobs. *Manual for the State-Trait Anxiety Inventory (Form Y1 – Y2)*, volume IV. Consulting Psychologists Press, 01 1983.

[SGT20] Prabha Sundaravadivel, Vividha Goyal, and Lakshman Tamil. i-rise: An iot-based semi-immersive affective monitoring

framework for anxiety disorders. In *2020 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–5, January 2020.

[SJS14] William Steptoe, Simon Julier, and Anthony Steed. Presence and discernability in conventional and non-photorealistic immersive augmented reality. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 213–218. IEEE, 2014.

[SKWL06] Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, 166(10):1092–1097, 2006.

[Sme10] Tom Smeets. Autonomic and hypothalamic-pituitary-adrenal stress resilience: Impact of cardiac vagal tone. *Biological Psychology*, 84(2):290–295, 2010.

[SMZ20] Fred Shaffer, Zachary M Meehan, and Christopher L Zerr. A critical review of ultra-short-term heart rate variability norms research. *Frontiers in neuroscience*, 14:594880, 2020.

[Sof19] Valve Software. Steamvr unity plugin. `https://valvesoftware.github.io/steamvr%5Funity%5Fplugin/`, 2019. Accessed: 2022-03-13.

[Spi13] Charles D Spielberger. *Anxiety and behavior*. Academic press, 2013.

[SR16] Susan H Spence and Ronald M Rapee. The etiology of social anxiety disorder: An evidence-based model. *Behaviour Research and Therapy*, 86:50–67, 2016.

[SS15] Murray B Stein and Jitender Sareen. Generalized anxiety disorder. *New England Journal of Medicine*, 373(21):2059–2068, 2015.

[SS22] Kristin L. Szuhany and Naomi M. Simon. Anxiety Disorders: A Review. *Jama*, 328(24):2431–2445, 12 2022.

[SSB+20] Dimitrios Saredakis, Ancret Szpak, Brandon Birckhead, Hannah AD Keage, Albert Rizzo, and Tobias Loetscher.

Factors associated with virtual reality sickness in head-mounted displays: a systematic review and meta-analysis. *Frontiers in human neuroscience*, 14:96, 2020.

[SSM16] Fredric Shaffer, Steven Shearman, and Zachary M Meehan. The promise of ultra-short-term (ust) heart rate variability measurements. *Biofeedback*, 44(4):229–233, 2016.

[SSMM98] Mel Slater, Anthony Steed, John McCarthy, and Francesco Maringelli. The influence of body movement on subjective presence in virtual environments. *Human factors*, 40(3):469–477, 1998.

[ST17] Federica Scarpina and Sofia Tagini. The stroop color and word test. *Frontiers in psychology*, 8:557, 2017.

[Str35] JR Stroop. Stroop color word test. *J Exp Physiol*, 18:643–62, 1935.

[TA+83] James T Townsend, F Gregory Ashby, et al. *Stochastic modeling of elementary psychological processes*. CUP Archive, 1983.

[TB06] Peter Tyrer and David Baldwin. Generalised anxiety disorder. *The Lancet*, 368(9553):2156–2166, 2006.

[TCT+15] G. Tartarisco, N. Carbonaro, A. Tonacci, G.M. Bernava, A. Arnao, G. Crifaci, P. Cipresso, G. Riva, A. Gaggioli, D. De Rossi, A. Tognetti, and G. Pioggia. Neuro-fuzzy physiological computing to assess stress levels in virtual reality therapy. *Interacting with Computers*, 27(5):521–533, 2015.

[Tea22] Audacity Team. Audacity. `https://www.audacityteam.org/`, 2022. Accessed: 2022-03-13.

[TGGC15] Lisa C. Thomas, Christopher Gast, Robert Grube, and Kimberly Craig. Fatigue detection in commercial flight operations: Results using physiological measures. *Procedia Manufacturing*, 3:2357–2364, 2015. 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015.

[Uni22a] Unity. Unity real-time development platform. `https://unity.com/`, 2022. Accessed: 2022-12-06.

[Uni22b] Unity3D. Build virtual worlds from scratch | unity probuilder, 2022.

[vDML19] J. Šalkevicius, R. Damaševičius, R. Maskeliunas, and I. Laukienė. Anxiety level recognition for virtual reality therapy system using physiological signals. *Electronics (Switzerland)*, 8(9), 2019.

[VK17] V Vanitha and Pandian Krishnan. Real time stress detection system based on eeg signals. *Biomedical Research 2016*, 2017.

[WC91] David Watson and Lee Anna Clark. Mood and anxiety symptom questionnaire. *Journal of Behavior Therapy and Experimental Psychiatry*, 1991.

[WC94] David Watson and Lee Anna Clark. The panas-x: Manual for the positive and negative affect schedule-expanded form. 1994.

[WCL+10] Dongrui Wu, Christopher G. Courtney, Brent J. Lance, Shrikanth S. Narayanan, Michael E. Dawson, Kelvin S. Oie, and Thomas D. Parsons. Optimal arousal identification and classification for affective computing using physiological signals: Virtual reality stroop task. *IEEE Transactions On Affective Computing*, 1(2):109–118, Jul-dec 2010.

[WCT88] David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

[Whe08] Robert Whelan. Effective analysis of reaction time data. *The psychological record*, 58(3):475–482, 2008.

[WKM19] Theresa F Wechsler, Franziska Kümpers, and Andreas Mühlberger. Inferiority or even superiority of virtual reality exposure therapy in phobias?–a systematic review and quantitative meta-analysis on randomized controlled trials specifically comparing the efficacy of virtual reality exposure to gold standard in vivo exposure in agoraphobia, specific phobia, and social phobia. *Frontiers in psychology*, page 1758, 2019.

[WKS+19] Siw Waffenschmidt, Marco Knelangen, Wiebke Sieben, Stefanie Bühn, and Dawid Pieper. Single screening versus conventional double screening for study selection in systematic reviews: A methodological systematic review. *BMC Medical Research Methodology*, 19(1):1–9, jun 2019.

[WL13] Guang-Xin Wang and Li Li. Virtual reality exposure therapy of anxiety disorders. *Advances in Psychological Science*, 20(8):1277–1286, Jul 2013.

[WMM96] J Mark G Williams, Andrew Mathews, and Colin MacLeod. The emotional stroop task and psychopathology. *Psychological bulletin*, 120(1):3, 1996.

[Wol90] Joseph Wolpe. *The practice of behavior therapy*. Pergamon press, 1990.

[WVDMG07] Eric-Jan Wagenmakers, Han LJ Van Der Maas, and Raoul PPP Grasman. An ez-diffusion model for response time and accuracy. *Psychonomic bulletin & review*, 14(1):3–22, 2007.

[WvVG+10] Klaas J Wardenaar, Tineke van Veen, Erik J Giltay, Edwin de Beurs, Brenda WJH Penninx, and Frans G Zitman. Development and validation of a 30-item short adaptation of the mood and anxiety symptoms questionnaire (masq). *Psychiatry research*, 179(1):101–106, 2010.

[WW98] Brenda K. Wiederhold and Mark D. Wiederhold. A review of virtual reality as a psychotherapeutic tool. *CyberPsychology & Behavior*, 1(1):45–52, 1998.

[YK19] Shengjie Yao and Gyoung Kim. The effects of immersion in a virtual reality game: Presence and physical activity. In *HCI in Games: First International Conference, HCI-Games 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings 21*, pages 234–242. Springer, 2019.

[YMW15] Gareth W Young, David Murphy, and Jeffrey Weeter. Auditory discrimination of pure and complex waveforms combined with vibrotactile feedback. *New Interfaces Music. Expr., Baton Rouge, LA, USA*, 2015.

[YS16] Miku Yanagimoto and Chika Sugimoto. Recognition of persisting emotional valence from eeg using convolutional neural networks. In *2016 IEEE 9th International Workshop on Computational Intelligence and Applications (IWCIA)*, pages 27–32. IEEE, 2016.

[ZTXS19] Lingqing Zhu, Xiang Tian, Xiangmin Xu, and Lin Shu. Design and evaluation of the mental relaxation vr scenes using forehead eeg features. In *2019 IEEE MTT-S International Microwave Biomedical Conference (IMBioC)*, volume 1, pages 1–4, May 2019.

[ZWLH16] Xu Zhang, Wanhui Wen, Guangyuan Liu, and Hui Hu. Recognition of public speaking anxiety on the recurrence quantification analysis of gsr signals. In *2016 Sixth International Conference on Information Science and Technology (ICIST)*, pages 533–538. IEEE, 2016.

Appendix A

PRISMA-ScR Checklist

Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
TITLE			
Title	1	Identify the report as a scoping review.	23
ABSTRACT			
Structured summary	2	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	23
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	23-24
Objectives	4	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	24-25
METHODS			
Protocol and registration	5	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	26
Eligibility criteria	6	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	25
Information sources*	7	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	25
Search	8	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	25-26
Selection of sources of evidence†	9	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	26
Data charting process‡	10	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	26
Data items	11	List and define all variables for which data were sought and any assumptions and simplifications made.	26
Critical appraisal of individual sources of evidence§	12	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate).	26
Synthesis of results	13	Describe the methods of handling and summarizing the data that were charted.	29

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
RESULTS			
Selection of sources of evidence	14	Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.	26-27
Characteristics of sources of evidence	15	For each source of evidence, present characteristics for which data were charted and provide the citations.	28-29
Critical appraisal within sources of evidence	16	If done, present data on critical appraisal of included sources of evidence (see item 12).	29
Results of individual sources of evidence	17	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	29-33
Synthesis of results	18	Summarize and/or present the charting results as they relate to the review questions and objectives.	29-33
DISCUSSION			
Summary of evidence	19	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	33-34
Limitations	20	Discuss the limitations of the scoping review process.	34-35
Conclusions	21	Provide a general interpretation of the results with respect to the review questions and objectives, as well as potential implications and/or next steps.	35-36
FUNDING			
Funding	22	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	Supplemental

JBI = Joanna Briggs Institute; PRISMA-ScR = Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews.
* Where *sources of evidence* (see second footnote) are compiled from, such as bibliographic databases, social media platforms, and Web sites.
† A more inclusive/heterogeneous term used to account for the different types of evidence or data sources (e.g., quantitative and/or qualitative research, expert opinion, and policy documents) that may be eligible in a scoping review as opposed to only studies. This is not to be confused with *information sources* (see first footnote).
‡ The frameworks by Arksey and O'Malley (6) and Levac and colleagues (7) and the JBI guidance (4, 5) refer to the process of data extraction in a scoping review as data charting.
§ The process of systematically examining research evidence to assess its validity, results, and relevance before using it to inform a decision. This term is used for items 12 and 19 instead of "risk of bias" (which is more applicable to systematic reviews of interventions) to include and acknowledge the various sources of evidence that may be used in a scoping review (e.g., quantitative and/or qualitative research, expert opinion, and policy document).

From: Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMAScR): Checklist and Explanation. Ann Intern Med. 2018;169:467–473. doi: 10.7326/M18-0850.

Appendix B

Extraction Form - Scoping Review

- Reference (First author / Year / Journal citation)
- Sample size
- Anxiety measurement/inducing method(s) used
- Number of outputs for classification
- Quality Score
- Highest Accuracy Achieved
- Notes (and additional bibliography)

Appendix C

Quality Assessment Scale

Supplementary Material

Quality Assessment Scale

- QS1:Was there a clear statement of the aims of the research?
  - Yes: Findings are explicitly stated and adequately discussed in relation to the initial research question
  - No: No information
  - Partially: The aims are stated but not sufficiently discussed
- QS2: Were the methods used justified in the paper?
  - Yes: The biosignals and machine learning methods used were described with the reasons and it had face validity
  - No: Not explained
  - Partially: Methods were explained but either not justified or not detailed enough and/or were open for interpretation
- QS3: Was the recruitment strategy appropriate to the aims of the research?
  - Yes: Sample size and population were determined before the research and justified as to why this population is the best for the aims of the study
  - No: There is no information regarding how the participants were selected
  - Partially: There was some information on the recruitment process but eligibility criteria
- QS4: Have ethical issues been taken into consideration?
  - Yes: The study received ethical approval from a committee and clearly states the policy number
  - No: No information
  - Partially: Ethical concerns were explained clearly for the reader to make out that enough was done to avoid them, but there was no mention of ethical approval from a committee

- QS5: Were the classification model(s) used clearly explained and sufficient?
  - Yes: There was a clear explanation of the analysis done and it was appropriate to study, specifically with diagrams and justification for them being used
  - No: No information
  - Partially: Process was briefly mentioned but without sufficient detail and/or open to interpretation
- QS6: Is there a clear statement of findings?
  - Yes: Findings are explicitly stated and adequately discussed in relation to the initial research question
  - No: No information
  - Partially: The findings are stated but not adequately discussed
- QS7: Are the limitations of the study thoroughly discussed?
  - Yes: There are references to sample and technical reasons that the study might not be generalisable
  - No: No limitations discussed
  - Partially: There are some limitations discussed, but there are some concerns that were not mentioned
- QS8: Are the contributions of the study to the literature and future research discussed?
  - Yes: There is an adequate discussion about the contribution of the study
  - No: No mention
  - Partially: Contributions to the literature are superficially discussed, with little evidence or justification from the findings of the paper and where they fit in the area

Appendix D

eStroop Code Snippets

```
if (stage < 3)
{
    breakTime = true;
    tooltipObj.SetActive(false);
    audioSource.Stop();
    WrapText.UnwrappedText = "Please take these " +
                             "15 seconds to relax " +
                             "before the next stage.";
    audioSource.PlayOneShot(audioClips[10]);
    WrapText.NeedsLayout = true;
    yield return new WaitForSeconds(15);
    WrapText.UnwrappedText = "The break is now over. " +
                             "Get ready for the next stage.";
    WrapText.NeedsLayout = true;
    audioSource.PlayOneShot(audioClips[14]);
    yield return new WaitForSeconds(3);
}
```

Snippet D.1: Code snippet for the break

```
var wordInd = 0;

PickColor();

switch (stage)
```

```
{
    case 1:
        wordInd = oldWord + 1;
        stroopText.text = neutralWords[wordInd].ToUpper();
        currentWord = stroopText.text;
        break;
    case 2:
        wordInd = oldWord + 1;
        stroopText.text = mildWords[wordInd].ToUpper();
        currentWord = stroopText.text;
        break;
    case 3:
        wordInd = oldWord + 1;
        stroopText.text = severeWords[wordInd].ToUpper();
        currentWord = stroopText.text;
        break;
}

StartCoroutine(nameof(CheckAnswer));

oldWord = wordInd;
```

Snippet D.2: Code snippet for task coroutine

```
var colorInd = 0;
colorInd = Random.Range(0, 4);
if (colorInd != prevColor)
{
    stroopText.color = colors[colorInd];
    prevColor = colorInd;
    correctAnswer = colorStr[colorInd];
}
else {PickColor();}
```

Snippet D.3: The pickColor function

```
var answerTime = 0f;
```

```
answerCorrect = false;
stroopClicked = false;

while (!(stroopClicked || answerTime >= 2f))
{
    answerTime += Time.deltaTime;
    yield return null;
}
if (stroopClicked)
{
    answerCorrect = (correctAnswer == stroopColor);
    Debug.Log(answerCorrect.ToString() +
              " Reaction Time: " +
              answerTime);
    logger.WriteEventToLog(answerCorrect.ToString() +
                           " Reaction Time: " +
                           answerTime);
    timesAnswered++;
    totalReactionTime += answerTime;
    if (answerCorrect)
    {
        totalCorrectAnswers++;
    }
} else
{
    Debug.Log("No Answer");
    logger.WriteEventToLog("No Answer");
}
```

Snippet D.4: countDown function code snippet

```
colorInd = Random.Range(0,4);
stroopText.color = colors[colorInd];

switch (stage)
{
    case 1:
        wordInd = colorInd;
```

```
        stroopText.text = colorStr[wordInd];
        break;
    case 2:
        wordInd = Random.Range(0,4)
        if (wordInd == colorInd)
        {
            StroopGen(stage);
            return;
        }
        break;
    case 3:
        wordInd = oldWord +1;
        stroopText.text = words[wordInd].toUpper();
        break;
}
```

Snippet D.5: Snippet of the previous Stroop task script

```
if (Input.GetButtonDown("Fire1")) {trigger = true;}
if (Input.GetButtonUp("Fire1")) {trigger = false;}

if (hover && trigger)
{
    Stroop.locked = false;
    transform.position = new Vector3(
                        activeHand.transform.position.x,
                        transform.position.y,
                        transform.position.z);

    if (transform.localPosition.x > 5)
    {
        transform.localPosition = new Vector3(
                            5,
                            transform.localPosition.y,
                            transform.localPosition.z);
    }
    else if (transform.localPosition.x < -5)
    {
```

```
        transform.localPosition = new Vector3(
                                -5,
                                transform.localPosition.y,
                                transform.localPosition.z);
    }

    anxiety = MapTo100(transform.localPosition.x);
}
```

Snippet D.6: VR Slider code snippet

Appendix E

Consent Form

Participant no: …..

Consent Form

I, …… hereby agree to participate in his study to examine user performance in a colour-word association task in VR conducted by Deniz Mevlevioglu under the supervision of Dr Sabin Tabirca and Mr David Murphy. □

I am 18 years old or older. □

I am participating voluntarily. □

The purpose of the study has been explained to me and I understand what is involved. □

I understand that I can withdraw from the study, without any repercussions. □

I understand that I can withdraw permission to use the data collected from me up to 14 days after my participation in the experiment, in which case the data will be deleted. □

I understand that my information will be kept confidential and my data will be anonymised. □

I give permission for my data to be used in any academic presentation or publication, including the PhD thesis. □

Signed Date

Appendix F

Pre-Questionnaire

Code: …..

Part 1. Answer the following questions.

Age group:

□ 18-24

□ 25-34

□ 35-44

□ 45-54

□ 55 or older

Gender:

□ Male

□ Female

□ Other

Have you experienced Virtual Reality before?

□ Never used

□ Only tried a couple of times

□ 1-2 years of experience

□ 2-5 years of experience

□ 5+ years experience

How often do you experience motion sickness?

□ Never

□ Rarely

□ Sometimes

□ Often

□ Very Often

Have you ever been diagnosed with colour blindness?

□ Yes

□ No

Have you ever been diagnosed with dyslexia?

□ Yes

□ No

Part 2. A number of statements people use to describe themselves are given below. Please read them and put an X in the box that best describes you *currently.*

	Not at all	**Somewhat**	**Moderately**	**Very Much**
I feel comfortable.				
I feel anguished.				
I feel at ease.				
I feel nervous.				
I feel concerned.				
Right now I feel good.				

Appendix G

Post-Questionnaire

Code: …..

Part 1. Please choose which option best describes how you felt during the different stages of the experiment.

Stage 1:

□ No anxiety □ Mild anxiety □ Severe anxiety

Stage 2:

□ No anxiety □ Mild anxiety □ Severe anxiety

Stage 3:

□ No anxiety □ Mild anxiety □ Severe anxiety

Part 2. Please read the following statements and mark the appropriate box representing to what extent the statement applies to you. If you're stuck between two options, choose the first one that came to your mind.

	Strongly Disagree	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
I found the equipment worn during the experiment cumbersome and uncomfortable.					
I was able to understand the instructions clearly.					
I was conscious of the devices on me at all times.					
Amount of equipment I was wearing reduced my immersion in the virtual world.					
I felt bored during the experiment.					
I was comfortable with the devices on me.					
I felt motion sick during the experiment.					
I found the platform with the buttons easy to use.					
I found the task confusing.					
The experiment had me engaged at all times.					
I wish the tutorial was more thorough.					

	Strongly Disagree	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
I found it difficult to understand and carry on the task given to me.					

Part 3. Is there anything you'd like to say to help improve the design of this experiment? (e.g. things to add or remove/change)

Appendix H

SPSS Variables

Table H.1: SPSS Variables

Name	**Measure**	**Label**
ParticipantCode	Nominal	Participant code
Agegroup	Nominal	Age Group
Gender	Nominal	Gender
Vrexposure	Ordinal	Exposure to Virtual Reality
MotionSickness	Ordinal	Motion Sickness
ColourBlindness	Nominal	Colour Blindness diagnosis
Dyslexia	Nominal	Dyslexia diagnosis
STAI	Scale	State-Trait Anxiety Inventory- Short version
GSR1	Scale	Average GSR in stage 1
GSR2	Scale	Average GSR in stage 2
GSR3	Scale	Average GSR in stage 3
Acc1	Scale	Accuracy in stage 1
Acc2	Scale	Accuracy in stage 2
Acc3	Scale	Accuracy in stage 3
RT1	Scale	Reaction time in stage 1
RT2	Scale	Reaction time in stage 2
RT3	Scale	Reaction time in stage 3
SelfVR1	Scale	Subjective units of Distress Scale after stage 1
SelfVR2	Scale	Subjective units of Distress Scale after stage 2
SelfVR3	Scale	Subjective units of Distress Scale after stage 3
Self1	Ordinal	Ordinal self-report anxiety in stage 1
Self2	Ordinal	Ordinal self-report anxiety in stage 2
Self3	Ordinal	Ordinal self-report anxiety in stage 3
SelfDiff	Scale	SUDS anxiety in stage 3 - stage 1
SCLpercent	Scale	Percentage increase in SCL from stage 1 to 3
OverallRT	Scale	Overall Reaction Time
ESE	Scale	Reaction time in stage 3 - stage 1
Q1	Ordinal	I found the equipment worn during the experiment cumbersome and uncomfortable.
Q2	Ordinal	I was able to understand the instructions clearly.
Q3	Ordinal	I was conscious of the devices on me at all times.
Q4	Ordinal	Amount of equipment I was wearing reduced my immersion in the virtual world.
Q5	Ordinal	I felt bored during the experiment.
Q6	Ordinal	I was comfortable with the devices on me.
Q7	Ordinal	I felt motion sick during the experiment.
Q8	Ordinal	I found the platform with the buttons easy to use.
Q9	Ordinal	I found the task confusing.
Q10	Ordinal	The experiment had me engaged at all times.
Q11	Ordinal	I wish the tutorial was more thorough.
Q12	Ordinal	I found it difficult to understand and carry on the task given to me.
Comments	Nominal	Open ended feedback about the experiment

Appendix I

Formal SPSS Results

Table I.1: Mauchly's Test of Sphericity

Measure	Mauchly's W	Approx. Chi-Square	df	Sig.	Greenhouse-Geisser
SCL	.123	56.620	2	<.001	.533
RT	.644	11.866	2	.003	.738
SUDS	.685	10.199	2	.006	.761

Table I.2: ANOVA Univariate

		Measure	df	F	Sig.
Condition	**SCL**	**Sphericity Assumed**	2	20.182	<.001
		Greenhouse-Geisser	1.065	20.182	<.001
	RT	**Sphericity Assumed**	2	1.996	.145
		Greenhouse-Geisser	1.475	1.996	.159
	SUDS	**Sphericity Assumed**	2	3.877	.026
		Greenhouse-Geisser	1.521	3.877	.039

Table I.3: Bonferroni pairwise comparisons of mean SCL between stages

Measure	(I) stage	(J) stage	Mean Difference (I-J)	SE	Sig.
SCL	1	2	-.579	.154	.002
		3	-1.000	.214	<.001
	2	1	.579	.154	.002
		3	-.421	.074	<.001
	3	1	1.000	.214	<.001
		3	-.421	.074	<.001

Appendix J

Framework Code Snippets

```
public void ApplyChanges()
{
    var prevLevel = expLevel;
    expLevel = expLevelDropdown.value + 1;
    waitingRoom.interactable = true;
    string sceneToLoad = expTypeDropdown.options[
                        expTypeDropdown.value].text;
    if (
    SceneManager.GetActiveScene().name != sceneToLoad
    )
    {
        SteamVR_LoadLevel.Begin(sceneToLoad);
    } else if (expLevel != prevLevel)
    {
        var target = GameObject.FindGameObjectsWithTag(
                            "exposure"
                            )[0];
        IExposure<int> targetScript = target.GetComponent(
                                    typeof(IExposure<int>)
                                    ) as IExposure<int>;
        targetScript.ChangeExposure(expLevel);
        Biosignals.lastChanged = Time.realtimeSinceStartup;
    }
}
```

Snippet J.1: Code snippet for exposure level management

Appendix K

Affective Words

neutral	**mildly emotional**	**severely emotional**
absurd	abduction	abuse
alimony	abortion	accident
aloof	ache	afraid
ankle	addict	agony
appliance	addicted	ambulance
bandage	alcoholic	anger
barrel	allergy	angry
bench	alone	annoy
bereavement	anguished	assault
bland	avalanche	bankrupt
blase	beggar	betray
board	blubber	blackmail
building	broken	bloody
bus	bullet	bomb
cabinet	burdened	brutal
cane	burial	burn
cell	carcass	cancer
cellar	cemetery	cockroach
chair	coffin	crash
chin	corpse	crucify
clock	corrupt	danger
coarse	crime	demon
column	criminal	despise
contents	crisis	destroy

neutral	mildly emotional	severely emotional
context	crude	detest
cord	cruel	devil
cork	crushed	disaster
corner	damage	disloyal
corridor	dead	distressed
curtains	death	divorce
dark	debt	dreadful
derelict	deceit	drown
detached	decompose	embarrassed
dirt	defeated	enraged
door	delayed	fear
dustpan	depressed	fearful
egg	depression	fraud
elbow	deserter	guillotine
engine	despairing	guilty
ennui	dirty	hate
errand	disappoint	hatred
finger	discouraged	horror
foot	disgusted	hostage
fork	displeased	hostile
fur	execution	humiliate
glass	failure	hurt
habit	fat	insane
hairdryer	filth	insult
hairpin	flabby	intruder
haphazard	foul	jealousy
hay	frigid	killer
headlight	frustrated	leprosy
hinder	funeral	lie
history	fungus	lost
hydrant	gangrene	mad
icebox	garbage	malice
indifferent	germs	murderer
inhabitant	greed	mutilate
ink	grief	nightmare
insect	gripe	pain

neutral	mildly emotional	severely emotional
iron	hardship	pervert
item	headache	poison
journal	hell	pollute
jug	helpless	punishment
kerchief	hooker	putrid
kerosene	illness	quarrel
kettle	immoral	rabies
knot	impotent	rage
lawn	infection	rape
lazy	injury	regretful
listless	insecure	rejected
locker	invader	riot
machine	jail	roach
mantel	lawsuit	rude
material	lice	scalding
meek	loneliness	scared
metal	lonely	seasick
mold	loser	sin
month	louse	sinful
muddy	maggot	slap
nonchalant	massacre	slaughter
nonsense	measles	slave
noose	menace	stress
nun	misery	suffocate
obey	mistake	surgery
odd	morbid	terrible
office	morgue	terrified
overcast	mosquito	terrorist
pamphlet	neglect	thief
paper	nuisance	tornado
part	obnoxious	torture
passage	offend	toxic
patent	paralysis	tragedy
patient	penalty	traitor
pencil	pest	trauma
phase	poverty	trouble

neutral	mildly emotional	severely emotional
pig	prison	troubled
plain	pus	tumor
pungent	rat	ulcer
radiator	robber	unfaithful
rain	rotten	upset
rattle	scandal	vandal
reserved	scar	venom
rusty	scorn	victim
seat	scornful	violent
serious	scum	vomit
shadow	scurvy	war
shy	selfish	whore
slow	severe	wicked
slush	shamed	wounds
solemn	sickness	
square	sissy	
stagnant	slime	
statue	slum	
stiff	smallpox	
stomach	snob	
stool	spider	
street	starving	
subdued	stupid	
table	suicide	
taxi	syphilis	
thermometer	tobacco	
timid	tomb	
tool	toothache	
trunk	ugly	
umbrella	useless	
utensil	wasp	
vest		
weary		