# Learning effective machine learning models for clinical applications in psychiatry

by

Jeffrey Sawalha

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Psychiatry

University of Alberta

# Abstract

This is a comprehensive examination of the diagnostic landscape in psychiatry and the role precision psychiatry might play in redefining how we classify illnesses. This thesis delves into three areas that currently exist in psychiatry today and is divided into five chapters. The first and second chapters review the literature and clinical space of AI and psychiatry. The last three chapters consist of academic articles that explore the use of AI in the three problematic areas of psychiatry.

The third chapter involves the early stage detection of Bipolar Disorder (Type 1) using cognitive assessments. Identifying cognitive dysfunction in the early stages of Bipolar Disorder (BD) can allow for early intervention. Previous studies have shown a strong correlation between cognitive dysfunction and the number of manic episodes. The objective of this study was to apply machine learning (ML) techniques on a battery of cognitive tests to identify first-episode BD patients (FE-BD). Specifically, we wanted to know if we could make generalized predictions about the various stages of BD using cognitive tests.

The fourth chapter examines childhood anxiety and produces a learned model that can detect dysfunction in the brains of children while they examine emotional facial expressions. Childhood anxiety is a difficult disorder to diagnose due to validity controversies and the conflation of normal developmental-behavioral patterns with anxiety symptoms. Our study not only seeks to train a model that can distinguish anxious from non-anxious children, but also to

discover neural markers related to this diagnosis.

Lastly, the fifth chapter utilizes natural language processing to detect the presence of post-traumatic stress disorder (PTSD). Specifically, we utilize sentiment analysis, a sub area of natural language processing (NLP), to extract emotional content from text information. In our study, we train an ML model on text data, which is part of the Audio/Visual Emotion Challenge and Workshop (AVEC-19) corpus, to identify individuals with PTSD using sentiment analysis from semi-structured interviews. We sought to understand the emotional spectrum of language and compare our findings with the ongoing literature.

Together, each of these studies illustrate how ML can be used to augment clinical decision-making surrounding the underlying conditions of individuals who may suffer from these illnesses. In doing so, we provide a conceptual review of the current barriers that exist in precision psychiatry today. Our hope is to provide the reader with a foundation of how ML can be used in psychiatry, while also highlighting some of the current barriers that hold back this field today. This comes in the form of a conceptual review (Chapter 2) and sets the landscape for the three published articles included.

# Preface

Some research conducted for this thesis forms part of an international research collaboration. All 3 papers are collaborations with different institutions across the globe. All secondary analyses received approval from the University of Alberta's Health Research Ethics Board (Pro 00072946).

First, the group at the University of Alberta includes Dr. Andrew Greenshaw, Dr. Russell Greiner, Dr. Bo Cao, Dr. Mohammed Yousefnezhad, Dr. Matthew Brown, Dr. Alessandro Selvitella (Purdue University) and Zehra Shah. All members played a pivotal role in one of the three papers. Dr. Greenshaw and Dr. Greiner were the main principal investigators for all content in this thesis.

Chapter 3 was a collaborative project with Dr. Tao Li from the Affiliated Brain Hospital of Guangzhou Medical University at Guangzhou Huiai Hospital in China. The data were collected and stored by Dr. Liping Cao. I conducted the main investigation, data analysis, writing and editing of the original manuscript.

Chapter 4 was a secondary analysis, using data from Dr. Kimberly Carpenter at Duke University. Tony Yousefnezhad and I did the primary investigation, data analysis, writing, and editing of the original manuscript. We also received approval from the University of Alberta's Health Research Ethics Board (Pro 00072946).

Chapter 5 was also a secondary analysis, using data from the University of Southern California (USC), and was approved by the USC Review board (UP-11-00342). Tony Yousefnezhad and I did the primary investigation, data analysis, writing, and editing of the original manuscript. Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian,

Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, Louis-Philippe Morency all collected and organized the data for this paper.

*To Ray McClellan*

*Thank you for believing in me.*

*All of the biggest technological inventions created by man - the airplane, the automobile, the computer - says little about his intelligence, but speaks volumes about his laziness.*

– Mark Kennedy

# Acknowledgements

I express my gratitude to the University of Alberta, IBM, and Alberta Innovates for providing financial support throughout my degree. I am humbled to have received grants and awards from these esteemed institutions.

I also extend my appreciation to my supervisors, Cloud, Russ, Andy, and Matt, for their unwavering faith in me, allowing me to grow both as a student and as an individual. I am thankful for their mentorship and guidance. I would also like to acknowledge Allen Chan, Glen Baker, Tara Checknita, and Suzette Bremault-Phillips for their personal mentorship and support. I could not have completed this journey without their assistance. Furthermore, I extend my gratitude to the two post-doctoral fellows, Alessandro and Tony, for their friendship and patience throughout my degree.

I am indebted to my parents, Aida and Shakib Sawalha, for providing me with opportunities to excel in my studies and career. This thesis is dedicated to both of them. I am particularly grateful to my father, who inspired me to persevere in my studies and career despite his own health struggles. He is my hero.

I am also thankful to my hockey team, the Edmonton Cobras, for keeping me grounded and providing me with stress relief during my studies. They were a significant source of happiness throughout my degree.

Finally, I would like to express my appreciation to the individuals who generously provided me with their datasets. I extend my gratitude to Kimberly L.H. Carpenter for allowing me to work with functional neuroimaging data of children suffering from anxiety. I also thank John Gratch and the team at the University of Southern California for providing me with access to the Distress Analysis Interview Corpus dataset. Finally, I extend my thanks to Bo Cao and

Tao Li for their collaboration regarding bipolar disorder and cognitive batteries. Their data and expertise were instrumental in this thesis.

# Contents


**1 Introduction** **1**

**2 Conceptual Review: Big data, big problems: Barriers that reside in precision psychiatry today** **8**
2.1 Data quality and inclusiveness . . . . . 10
2.1.1 Data Ethics and Privacy . . . . . 12
2.1.2 Validity of diagnostic and prognostic labels . . . . . 14
2.1.3 Machine Learning, explainability and human intuition . 15
2.1.4 Conclusion . . . . . 18

**3 Individualized identification of first-episode bipolar disorder using machine learning and cognitive tests** **20**
3.1 Introduction . . . . . 20
3.2 Methods . . . . . 23
3.2.1 Participants . . . . . 23
3.2.2 Cognitive Assessment . . . . . 24
3.2.3 Data Cleaning . . . . . 24
3.2.4 Machine Learning Analysis . . . . . 26
3.2.5 Statistical Analysis . . . . . 27
3.3 Results . . . . . 27
3.3.1 Group Analysis . . . . . 27
3.3.2 Machine Learning . . . . . 27
3.4 Discussion . . . . . 29
3.4.1 Supplementary Tables . . . . . 36

**4 Predicting pediatric anxiety from the temporal pole using neural responses to emotional faces** **39**
4.1 Introduction . . . . . 39
4.2 Methods . . . . . 43
4.2.1 Data and Code . . . . . 43
4.2.2 Recruitment . . . . . 44
4.2.3 Participants . . . . . 45
4.2.4 Functional MRI task . . . . . 45
4.2.5 MRI acquisition . . . . . 46
4.2.6 Pre-processing . . . . . 47
4.2.7 Analysis: Anxious versus non-anxious classification . . 48
4.2.8 Analysis: Negative facial stimuli . . . . . 51
4.3 Results . . . . . 53
4.3.1 Clinical and demographic statistical analysis . . . . . 53
4.3.2 Anxious versus non-anxious classification analysis . . . 53
4.3.3 Negative stimuli classification . . . . . 57
4.4 Discussion . . . . . 60
4.4.1 Neuroanatomy of the Temporal Pole (TP) . . . . . 62

4.4.2 Evidence of socioemotional processing in the temporal pole 63
4.4.3 Implications in Childhood anxiety . . . . . 64
4.4.4 Limitations and future work . . . . . 68
4.5 Conclusion . . . . . 69

**5 Detecting presence of PTSD using sentiment analysis from text data 72**
5.1 Introduction . . . . . 72
5.2 Related Work . . . . . 75
5.3 Methods . . . . . 78
5.3.1 Participants . . . . . 78
5.3.2 Procedure . . . . . 79
5.3.3 Transcription . . . . . 81
5.3.4 Preprocessing . . . . . 81
5.3.5 Sentiment analysis . . . . . 82
5.3.6 Machine learning . . . . . 82
5.4 Results . . . . . 86
5.4.1 Demographics . . . . . 86
5.4.2 Machine Learning and Statistical Analysis . . . . . 88
5.4.3 Bin Analysis . . . . . 89
5.5 Discussion . . . . . 91

**6 Conclusion 99**

**References 105**

# List of Tables


3.1 Clinical and demographic descriptions of two cohorts . . . . . 25
3.2 Cogntive outcome measures . . . . . 37
3.3 List of model beta weights from linear SVM. . . . . 38

4.1 A summary table of demographic and clinical symptom scores of participants . . . . . 58
4.2 Talairach regions most correlated with region #41 . . . . . 59

5.1 Demographics and outcome measures for AVEC-19 participants 86
5.2 Demographics and outcome measures for oiriginal partitioning folds . . . . . 87
5.3 Demographics and outcome measures for 2017-to-2019 partitioning folds . . . . . 87
5.4 Performance of models across sentiment analyzers and partitioning schemes . . . . . 90
5.5 Average binned sentiments per group . . . . . 92

# List of Figures


3.1 Confusion matrix of model results for the 2 cohort datasets . . 28
3.2 Top predictors from Linear SVM . . . . . 29
3.3 Receiver operating characteristic curve (ROC) . . . . . 30
3.4 Model prediction (FE) grouped by mood state . . . . . 31
3.5 Model prediction (CHR-BD) vs. HC . . . . . 32

4.1 Processing pipeline for selecting the best Talairach region . . . 42
4.2 Top 20 Talairach regions by accuracy (%) . . . . . 55
4.3 Individual accuracies (%) of participants based on sub-classes from final super learner model . . . . . 55
4.4 Individual accuracies grouped across age, group, and scanner site 56
4.5 Grouped Bayesian representational similarity analysis of region #41 . . . . . 58
4.6 Posthoc analysis: Preprocessing pipeline for negative stimuli analysis . . . . . 60
4.7 Precision tables for both anxious versus non-anxious and negative stimuli classification analysis . . . . . 71

5.1 Interview process with Ellie . . . . . 77
5.2 Sentiment analysis pipeline . . . . . 80
5.3 Machine learning results (F1 score) from partitioned types . . 91
5.4 Binned sentiment scores per group . . . . . 93

# Chapter 1

# Introduction

Since its inception, psychiatry has been hindered by tropes not otherwise seen in other medical fields. In cardiovascular studies, a heart attack is characterized by universal biomarkers, and it is accompanied by objective diagnostic tests [1]. Screening for cancer involves conducting biopsies, laboratory tests and imaging procedures to identify malignant tumors with high precision and sensitivity [2], [3]. In medical fields such as these, any two doctors, presented with the same patient, would be able to agree on what diagnosis is present, and that the diagnosis would mean the same thing to every other physician. In psychiatry, reliability among diagnostic categories is not trivial. Advancements towards accurate psychiatric diagnosis and treatment have been sporadic throughout history, jumping between different major paradigms. From Freud's psychoanalytic approach to diagnosing mental illness, to the biological pathology of Kraepelins' work, to the biopsychosocial model proposed by George Engel, psychiatry has been chasing a moving target, trying to find a valid diagnostic criterion to establish meaningful treatments and conduct universal scientific research [4], [5]. These abrupt paradigm shifts have led to an identity crisis in psychiatric practice, whereby the lack of direction has stymied the field.

The primary manual used to classify these disorders, the Diagnostic and Statistical Manual of Mental Disorders (DSM) [6], has undergone major shifts through its iterations. The way we classify mental illnesses are seen through the observable signs and symptoms gathered from this manual. Unfortunately, this has led to an unstable foundation by which psychiatry has been built upon.

Diagnoses still remain uncertain, as kappa values are low for several major psychiatric illnesses [7], [8]. Large comorbidities exist within the silos of these diagnostic categories, prognostic outlooks are unclear in psychotic illnesses (such as schizophrenia) [8]–[10], and therapeutic interventions have proven to be effective in only 30-50% of patients [11]–[13].

Steve Hyman, former director of the National Institute of Mental Health (NIMH), stated that the DSM "created an unintended epistemic prison that was palpably impeding scientific progress... Even animal studies that purported to develop disease models... were often judged by how closely they approximated DSM disorders" [14]. He claims that diagnostic labels only serve as placeholders, that can only provide agreement among psychiatrists until further objective measures (ones that have a biological underpinning) can be developed [15]. According to him, the DSM was developed to provide a common language based on observable signs and symptoms, but is explicitly devoid of pathophysiology or treatment response [16]. Building a diagnostic system like this may never yield the specificity needed to match those in other medical fields. Beyond that, the heterogeneity of symptoms across mental illness adds to this complication. As a result, diagnosing and treatment may be limited to the relief of presented symptoms, but do not target the underlying etiological factors of an illness such as depression [16]. As you will see in the next section, morbidity rates, disability adjusted life years (DALY), and costs associated with psychiatric illnesses have been on the rise.

Mortality rates for medical conditions such as heart disease, Leukemia, H.I.V, and stroke have seen a steady decline since the 1960s, thanks to scientific advancements [17]. For example, today, if detected early enough (within 3 hours), 30% of all patients who have a stroke will be discharged from the hospital, with no outstanding symptoms [17]. However, mental health has not seen the same type of progress. Compared to these other medical conditions, mental health has shown an inverse pattern since the 1960s. Globally, 1 in 5 people will now meet the criteria for a mental illness at some point in their lives. Suicide is now the 3rd highest cause of death among 15-24 year olds in North America [18]. Depression affects almost 300 million people worldwide per year

[19]. In the United States, almost 10% of all children (ages 2-17) will receive a diagnosis of attention deficit hyperactivity disorder (ADHD), 7% will receive an anxiety diagnosis, and 1 in every 6 children (ages 2-8) will be diagnosed with a mental, behavioral, or developmental disorder [20]. When examining DALYs, which is a metric for burden of disease, neuropsychiatric illnesses are the 5th highest medical condition worldwide [21], [22]. In Canada, it is the leading cause of disease burden [23]. In the United States, major depression alone is the 7th highest cause of DALY. Over the last 20 years, there has been far greater progress in reducing the burden of some major medical conditions, such as ischemic heart disease and lung cancer. However, little progress has been made for mental health disorders [21].

When accounting for the financial burden of psychiatric inpatients, their costs outweigh those in other major medical categories [24]. In Canada, high-cost psychiatric patients (a subset of patients who accrue the majority of health care costs) have an average health-care cost of $31,611, compared to other high-cost patients, who average $23,681 [24]. In 2011, mental health problems and illnesses cost the Canadian economy at least $50 billion [25]. $42 billion dollars were spent on psychiatric services in 2011, and that figure has only grown since. From that source, it was estimated that in 2021, mental health services would cost a national $80 billion dollars, and in 2031, $156 billion dollars [25]. Keep in mind that those projections do not account for the current COVID-19 global pandemic. These statistics prove that our strategy for treating mental illness is not optimized globally, nor are we containing the costs associated with it. Luckily, new approaches and technologies are paving the way to improve how we treat mental illness.

Hyman and his eventual successor, Thomas Insel, sought to create an alternative approach to classifying mental disorders that would begin with, but not be limited to, symptoms [16]. In 2009, the Research Domain Criteria project (RDOC) was created with the ultimate goal of precision medicine for psychiatry. Precision medicine is a new branch of medicine that uses several clinic-pathological laboratory measures (genetic tests, blood samples, medical history, and demographic factors) to tailor treatments to individuals instead of

large clinical groups [26]. RDOC is a criterion that attempts to develop a deeper understanding of the biological, genetic, and psychosocial factors for psychiatric illnesses, which may assist in how we treat these diseases [16]. Though it is not currently being used as a diagnostic manual, RDOC is used to: 1) organize research, 2) collect vast amounts of data — from molecular factors to social determinants — to attempt to understand abnormal behavior [16], and 3) use data to anchor diagnostic classification in a scientifically supported model [26]. Some notable marker discoveries include the Dexamethasone Suppression Test (DST) (an indicator of cortisol reactivity), to detect illnesses associated with a dysregulated sympathetic nervous system [27], the discovery of reduced rapid eye movement (REM) latency as a marker for major depression [28], and the use of smooth pursuit eye movements dysfunction to detect schizophrenia [26], [29]. Though these methods have not been perfect, they do point to a future where objective measures from the various domains of the RDOC may serve to detect and diagnose illnesses.

With the extensive broadening of functional domains and units of analysis from RDOC, a vast amount of data is being collected, and analyzed. However, traditional statistical techniques such as hypothesis testing, power analysis, and effect size measurements, which compares clinical groups with a hypothesized population, have come under criticism. By randomly sampling, inferences could be drawn about observable, quantifiable differences between the groups, without those changes unlikely being due to chance [8], [30]. This approach, known as the classical inference paradigm, served psychiatry and psychology for the better part of the $20^{\text{th}}$ century [8]. However, two major criticisms have been brought to light: 1) replication and reproducibility of previous studies have led to a crisis, whereby replication rates are as low as 11% for preclinical studies, and empirical estimates indicate that there is a 70% chance that significant results are false positives, even with multiple comparison corrections [31]–[33], and 2) group-based analysis ignores individuals differences, which results in validity issues when applying translational care. Often times, latent factors may jeopardize the inferences between and within groups, ultimately affecting the generalizability of these traditional approaches. Thus, clinical

translation to individuals is negatively affected, and they are left to suffer for years without truly knowing their underlying condition and do not receive adequate treatment.

Machine learning (ML) is a branch of artificial intelligence that analyzes large amounts of data to make inferences between covariates and outcome labels, even if non-linear relationships exist. This is done through two major methods, supervised and unsupervised learning. In supervised learning (the primary focus of this thesis), a model learns from a dataset with labelled instances (training). Once it has learned from the training phase, it is evaluated against an unseen dataset to make predictions about those labelled instances. In the 1950s, The perceptron was one of the first models used to that detect patterns of light and shade representing various shapes (e.g., squares and triangles) from photos of light sensors [34]. After seeing many labeled instances, the machine would modify the weights associated with each light sensor using an objective function. This was optimized to minimize the error between predictions and real labels of the instances. Misclassifications led to an adjustment of those statistical weights. Once trained, the perceptron made autonomous predictions about unlabeled instances, which were evaluated by human users [8]. In the same way, ML has been used to learn patterns of behavior, cognition, and biological processes in individuals with and without mental illnesses. The eventual hope is that these algorithms can make accurate predictions about individuals with psychiatric conditions, given that they are trained on a large, diverse dataset available. Notably, these algorithms are not to supplant human intuition, but to augment their decision-making process about the nature of a patients' illness, and how to cater treatment for the individual, not the group. After all, the labelled instances that are used in supervised learning come from human evaluation, which can be seen as both a limitation and an advantage. These points will be further discussed in the conceptual review for this thesis.

There are a multitude of areas that ML can integrate with psychiatry, but my thesis focuses on 3 nuanced areas that relate to differential diagnosis. The goal of my thesis is to illustrate that ML can be used to enhance the accuracy of diagnosis in the earlier stages of a mental disorder. This is not a trivial

problem because patients are often misdiagnosed due to human error, or a misunderstanding of symptomatology, which leads to more personal suffering, additional medical costs, and longer prognostic outlooks [35]. Accurately identifying the condition of a patient is the first step to preventing disease, as it is in other major medical fields. With the birth of the RDOC, now is the time to examine all biological and social functions within the vast categories of mental diseases, to look for clues that may uncover etiological factors. More objective markers are needed to consolidate the cluster of symptoms that patients present upon initial assessment. I believe that the classification of mental illnesses has reached a saturation point, as human intuition can only take us so far. A phenomenological approach, where we collect vast amounts of data, use ML to examine the patterns, and use inductive reasoning to form hypotheses should be considered going forward. This may be an optimal strategy to incorporate the diagnostic labels from the DSM with the input/predictors of the RDOC. The mission statement of the NIMH states that "the road to better therapeutics starts with better diagnosis", and this thesis echoes that sentiment. This thesis serves to illustrate how ML can validate and progress diagnostics in psychiatry, which could ultimately help treat individuals.

Outlined below are three salient issues within the field of psychiatry that may benefit from ML. These three issues make up the majority of this thesis. The thesis was mainly written by me, but several individuals contributed to the three publishes articles below. The first study deals with identifying individuals with bipolar disorder (BD) (Type I), using cognitive assessments. The overall research question here asks whether we can detect subtle abnormalities in cognitive functioning, for those who have experienced early signs of BD disorder. Liping Cao, Jianshan Chen, Alessandro Selvitella, Yang Liu, Chanjuan Yang, Xuan Li, Xiaofei Zhang, Jiaqi Sun, Yamin Zhang, Liansheng Zhao, Liqian Cui, Yizhi Zhang, Jie Sui, Russell Greiner, Xin-Min Li, Andrew Greenshaw, Tao Li, and Bo Cao are names of individuals involved in this study. My specific role was analyzing the data, training the machine learning model, and writing the manuscript.

The second examines brain states in children suffering from various types

of anxiety disorders, while they view angry and fearful faces in a magnetic resonance imaging (MRI) machine. In this study, we wanted to determine whether we can use neural correlates to identify which children suffer from several types of anxiety disorders. The members that contributed to this paper were Tony Muhammad Yousefnezhad, Alessandro M Selvitella, Bo Cao, Andrew J Greenshaw, and Russell Greiner. Tony and I conducted the analysis, and wrote the manuscript together. Dr. Greenshaw, Dr. Cao and Dr. Greiner edited and oversaw the study. Kimberly Carpenter was responsible for providing the data online, and conducted the original study.

The third paper examines detecting PTSD in military veterans using textual data from semi-structured clinical interviews. Here, we examined whether we could use the emotional content from transcribed clinical interviews to determine whether someone is suffering from PTSD. Our goal was to examine the emotional spectrum of language for individuals suffering from PTSD. Tony Muhammad Yousefnezhad, Zehra Shah, Matthew R. G. Brown, Andrew J. Greenshaw, and Russell Greiner were all part of this study. Specifically, Tony and I wrote the manuscript, Zehra conducted the literature search, Dr. Brown, Dr. Greenshaw and Dr. Greiner oversaw the project, and provided feedback on analysis and manuscripts.

The present thesis seeks to address the research inquiry of whether predictive models are a viable tool in a clinical context for strengthening and substantiating actionable determinations regarding diagnosis and prognosis. Further, we discuss the feasibility and practicality of such tools in the current healthcare environment. Each section comes with a publication, outlining the performance task, the current issue in psychiatry, the justification for ML, results and analysis. The novelty of this thesis illustrates that various types of data can benefit some of these nuanced problems in psychiatry. Additionally, since each paper provides a literature review on the respective topic, a conceptual review is presented. This conceptual review explores the current barriers in precision psychiatry today, and where we are headed in the future. The hope is that ML can provide additional solutions to diagnostics and therapeutics in psychiatry, with a weighted emphasis on the diagnostics segment.

# Chapter 2

# Conceptual Review: Big data, big problems: Barriers that reside in precision psychiatry today

In the introduction, I spoke about the ever-changing landscape of psychiatry and how it has led to a suboptimal solution for treating mental illness. I then introduced machine learning as a potential tool that could help us further detect, understand and treat psychiatric illnesses in some capacity. In practice, this is easier said than done. Today, big data (labeled data) is becoming ubiquitous, and algorithms are able to intake vast amounts of multi-modal data, to detect patterns and make predictions about the future with the variables at hand. This is especially true in other fields, such as business, agriculture and technology. While there is unrealized potential, computational psychiatry is still in its embryonic stages. Some major questions loom, such as: What systematic changes need to happen for us to capture the true potential of precision psychiatry? What kind of data should we be looking to collect? Do we currently have the right infrastructure to wrangle big data for psychiatry? What about data privacy and ethics? What about the implementation of these learned models? What if they outperform human expertise based on their performance evaluations (which are usually human constructed)? Can we interpret these models? Can ML become the primary decision-making source in a clinical setting instead of humans?

This conceptual review will be covering some of the major questions that surround computational psychiatry today. Working as a PhD student for the past 4 years, I have experiences that allow me to discuss these questions, as they were present in some of my own thesis projects. The overall goal of this review is to shed light on some prominent issues in translational psychiatry, and to provide an outlook on how psychiatry may look in 10 years, if we attempt to address these issues. The primary focus will be centered around 4 major topics: (1) Data quality and accessibility, (2) Data ethics and privacy, (3) The validity of diagnostic and prognostic labels and 4) Machine learning, explainability and human intuition. This conceptual review should provide the reader with critical questions, as well as solutions, surrounding the use of ML in psychiatry.

Two major roles can be considered for precision psychiatry. The first involves training models that can detect or categorize individuals in the prodromal stages of a certain illness. The idea that — identifying illnesses in the prodromal stage can improve prognostic outlooks [16]. This usually involves training on large amounts of labeled patient data (many characteristics – blood samples, neuroimaging, behavioral assessments, and clinical scales – from many subjects: patients and controls) to produce a model that can diagnose a patient or predict the effectiveness of a treatment. Once learned, a model can predict the clinical outcomes for new patients who may be exhibiting symptoms [36]. This is important because early detection of a psychiatric illness (such as Bipolar Disorder) usually leads to a better prognostic outlook [37]. The second role involves tailoring individualized interventions. For example, patient 1, who has been diagnosed with major depressive disorder (MDD) has been dealing with thoughts of suicide, insomnia, anhedonia, and asociality. Patient 2 has also been diagnosed with MDD, but she is suffering from hypersomnia, alcohol addiction, psychomotor agitation and weight loss. Though both fall under the category of MDD, they are very different, meaning the same treatment may not render the same results. Additionally, patients 1 and 2 may come from vastly different backgrounds, may have grown up in different cities and cultures, and have different medical histories. Instead of a trial-and-error approach to finding the right treatment for patient 1 and 2, it would be better if physicians could

recommend which type of antidepressant may be most effective for each, based on a computational model, which have been trained on thousands of samples (from electronic health records). In both scenarios, the clinical deployment and workflow of these ideas need to be first established. Several barriers stand between utilizing these ML models in a clinical setting.

## 2.1 Data quality and inclusiveness

George Fuechsel, an early IBM programmer and instructor, said the famous line when discussing data quality for ML: "Garbage in, garbage out". Data quality and accessibility in psychiatry has not yet been fully realized. There are several reasons for this. First, the way we collect data is not standardized, nor does it always provide the information needed to produce models that can make accurate predictions. Technology has advanced at an exponential rate over the last 20 years, but information infrastructure (from hospitals and clinics) have struggled to keep pace. For example, Alberta only recently digitized and reconstructed their electronic health records under the banner of Connect Care, to improve how they store their data [38]. Prior to this, data structures were inconsistent, with many missing variables, a lack of standardized records and reporting, and laborious complexity when it came to extracting data in a concise and organized manner. Today, healthcare systems like Alberta Health Services have moved towards a more ML friendly data infrastructure. This may be the ideal opportunity to achieve new global standards for how we standardize and assure data quality.

Increasing the volume of data can achieve better performance for learned models. Giving a learned model, more data; (collected from more locations), typically produces more robust and generalizable models. One of the major challenges in psychiatry is the need for standardized data across different hospitals and clinics to capture patterns that can identify and treat psychiatric illnesses. Most institutions measure and collect their own set of variables, which makes it more difficult to compare datasets or models to one another [39]. Ideally, a global initiative would have the various hospitals, clinicians and

researchers agree on a standardized version of data collection. That way, a transfer of knowledge can be shared when trained models have seen the same data, but from different locations. Luckily, I have been privileged to work with the HiMARC (Heroes in Mind, Advocacy and Research Consortium) research group, on a project that involves predicting treatment outcomes for military veterans who suffer from PTSD. They are conducting the same study in 6 different locations worldwide (Netherlands, the United States (California and Maryland), the United Kingdom (Wales), and Canada (Alberta and Ontario)) [40], collecting the same type of data at all 6 locations, with the goal of producing a model that could accurately predict how well a patient, from any of these locations, will respond to a specific therapy. This is a wonderful microcosm of what collaborative and global strategic planning can do for data quality and standardization in psychiatry.

Although data standardization is important, data quality and inclusiveness is equally vital. Algorithmic bias is a phenomenon that occurs when a learned model produces systematically prejudiced results, due to the underlying biases of the data it is trained on [41]. Thus, health data records contain implicit biases (such as under-represented populations, due to access of resources) that are not addressed when the data is collected. How and where we collect data can lead to algorithmic biases in mental health too [42]. These implicit biases usually operate outside of working consciousness, but they become clear when examining the output made by models, perhaps due to disproportionate representation of certain populations who have received clinical care. These biases can ultimately reflect in trained models, which are trained on large amounts of skewed patient data. As a famous example, Coley *et al.* 2021, examined racial/ethnic disparities in the performance of a suicide risk algorithm trained on health record data (of various ethnicities and various sample sizes) to predict suicide death in the 90 days after an outpatient mental health visit. Area under curve (AUC) scores showed a much higher sensitivity for Caucasian, Hispanic and Asian populations, but showed much lower sensitivities for African and Native American ethnicities. The authors concluded that the model was inaccurate in predicting suicide amongst African and Native Americans, and

these predictions could be potentially harmful for capturing suicidal risk [43]. They posited whether several models could be utilized to account for different subpopulations. Several latent factors affect the disproportionate representation of a clinical population. Factors such as socioeconomic status, culture, political reforms and policies, geographic location and genetics can still influence a model, even if these variables are not accounted for in a training dataset. Ensuring data quality requires precision regarding the population to which it applies.

### 2.1.1 Data Ethics and Privacy

One performance task of computational psychiatry involves discovering early signs of mental illness. If successful, this could improve outcomes and prevent future disability [39]. But what happens when these same analytics can reveal unwanted or unfavorable outcomes regarding a person's mental state? For example, in one of my previous projects, our goal was to create a computational tool (using speech data) that could determine who is at risk of developing PTSD, within the Canadian military. Imagine an individual who takes this screening test, and is told he/she is at high risk of developing PTSD, therefore, they cannot enlist in the army [39]. These types of scenarios may negatively affect the quality of life for individuals, and it may shut the door on certain possibilities because they are predisposed to a certain illness. Is it incumbent upon institutions and governments to abide by the declarations of the model? Or should they forgo the risk to allow freedom of choice for the individual? Who should have access to this decision? Should the person be informed about their failed screening test? Should they be given an option to find out? How will this affect the individual's quality of life if they know they are developing an illness? What if the screening accuracy is not 100%? These types of ethical dilemma may permeate the deployment of predictive analytics in psychiatry. Policymakers are still lagging behind with respect to these questions, but these technological advancements are forcing us to consider these ethical dilemmas and whether they infringe on our civil, social, and legal freedoms [39].

The second major barrier is data privacy and ownership. Today, humans

generate data at a breathtaking pace: each person generates an average of 1.67 megabytes per minute, which means the world generates approximately 1.145 trillion MB per day [44]. A large part of human data contains sensitive information about a person's current physical and mental conditions – either explicitly or implicitly. With the current surge in web-based apps that use biometric, facial, motion-tracking and text data – intimate parts of our identities – are given to corporations that are learning more than we can ever learn about ourselves. This data is being used to predict how an individual behaves in future circumstances [45]. Companies that have this data may be able to produce personalized or targeted advertising, or they may predict behaviors such as violence or suicide, or they may recommend certain articles based on political leanings. In a way, humans have become transformed products for large technological corporations, and their data is becoming as valuable as currency. The question remains: is it permissible to aggregate multiple streams of data to invade the privacy of members in our society, for the sake of monitoring and preemptive interventions? In an ideal world, participants would have the right to know exactly where and what their data is being used for, and more importantly, be able to direct it for certain uses. Informed consent should be a refined ethical requirement, whereby the participant knows the answers to these questions [46], [47]. Researchers or corporations should thoroughly inform users about risks and benefits of data use, and provide a detailed explanation of what they are doing with the data. Mobile applications or websites could institute comprehension assessments in the form of quizzes or games to clarify the user's understanding of their own data usage [48], [49]. To protect against the unintended consequences of data privacy and ownership, governments, researchers and corporations should adopt a novel, comprehensive guide to how they plan to use the data, who they plan to share it with (if consent is agreed upon), and where it would be stored [47]. This participant-centric model may help give agency back to those who have had their data wrangled without truly knowing the consequences of their decisions.

### 2.1.2 Validity of diagnostic and prognostic labels

As mentioned in the introduction, psychiatry has struggled with objective labels due to a lack of biological basis. Thus, when using ML to predict the presence of an illness, how can we be certain about the target label? What is ground truth in scenarios such as these? Is it an evidence-based decision from one clinician or many? From a practical perspective, For example, one of the projects in my thesis examined PTSD for individuals who partook in an interview [50]. For that dataset, diagnostic labels were decided based on a self-report questionnaire, the Post Traumatic Stress Disorder (PCL-C). This questionnaire is not the gold-standard for diagnosing PTSD, and so the validity of the labels were called into question, which ultimately would affect the reliability and validity of our trained model. But even if trained psychiatrists conducted gold-standard tests, disagreements about diagnosis would still exist [8]. For some disorders, the kappa values, which is a metric to determine reliability amongst clinicians, falls well below 0.4 [51]. The reliability of psychiatric diagnoses remains a significant challenge in the field. In particular, low reliability measures have been attributed to a variety of factors, including patient inconsistency (5%), clinician inconsistency (32.5%), and limitations in the nomenclature used in the Diagnostic and Statistical Manual of Mental Disorders (DSM) (62.5%) [52]–[54]. The development of computational tools based on such noisy ground truth labels raises concerns regarding the validity and reliability of resulting models.

Despite this issue, several studies have utilized diagnostic labels generated by physicians, although the ground truth labels for two out of three studies were determined by aggregating opinions from multiple physicians to reach a consensus. However, it is important to note that reliable and objective diagnostic labels are currently obtained primarily from trusted medical sources within our healthcare system, and this dependence on expert opinion poses a challenge for the development of more reliable diagnostic tools [52].

Compounding the issue is the limited and reductionist approach to psychiatric illness, which assumes that all diseases originate solely from the brain [52].

To address this, it has been suggested that a more holistic approach should be taken, with the inclusion of social, emotional, and idiosyncratic data about the individual, in addition to biological data [52].

In light of these challenges, it is crucial for future research to consider alternative methods for establishing more reliable and valid diagnostic tools. This may involve exploring new sources of ground truth data, as well as adopting a more comprehensive approach to data collection and analysis. Some suggestions have offered a holistic approach to the type of data we collect. Specifically, instead of solely focusing on biological data, it may be useful to also include social, emotional and idiosyncratic forms of information about the individual [52]. Additionally, rather than using summary measures and broad diagnostic categories, future ML research could identify very specific symptoms or labels, such as cognitive deficits in executive function or language, or even biological markers that span multiple diagnostic boundaries [8], [55]–[57]. Refining the labels in which we train our data on may reduce noise, and offer incremental, but more accurate insights into the ailing factors of a mental illness. Although the ultimate goal is to relieve symptoms of mental illness, properly identifying labels will allow for healthcare systems to approach treatment in a standardized, efficient way.

### 2.1.3 Machine Learning, explainability and human intuition

One final barrier remains: How should a learned model be used in clinical practice? There is a common rhetoric that ML will supplant human intuition when it comes to clinical decision-making [58]. This is somewhat of a distorted view of how ML can augment clinical decision-making. Firstly, humans are currently the ones who predict clinical outcomes and make informed decisions about a patient. But humans are not always accurate, and some decisions lead to better outcomes than others. Machine learning can assist in optimizing clinical decisions that will lead to the best outcomes [59]. Learned models can aggregate the thousands of observations made by humans to learn about favorable outcomes, and which variables affect those outcomes. They simply

select the best fitting statistical relationship between the features and the outcomes. For example, one of my thesis projects involves predicting pediatric anxiety using a task based functional neuroimaging task, where children view images of angry and fearful faces. In this study, terabytes of brain images are scanned by a functional neuroimaging machine, and the outcome labels are determined by a series of clinical assessments, done by a psychiatrist. Our model simply learns a function that can distinguish anxious from non-anxious children by measuring brain activity. Then, our learned model makes predictions about future children, and those predictions are compared to the diagnostic labels derived from human intuition. At every turn, our model compares its own performance with human outcomes. Upon examining this trope closer, it is apparent that ML is here to augment clinical decision-making, not replace it. However, even if that did happen, a larger problem looms — How do these models come to make decisions? And can these decisions be explained and trusted by humans?

Over the next 10 years, ML algorithms may play a more prominent hand in screening for psychiatric illnesses or personalizing treatment for an individual. If so, these decisions will have real world consequences, and human lives will be impacted based on those predictions. Therefore, clinicians and patients must be able to trust the output of these models, and come to understand how these models make the decisions that they do. This is part of a new, growing field called "explainable AI". The word "explainable" does not seem to have a rigorous definition yet [60], [61], but some have defined it as the "capability of a subject matter to be faithfully translated into a language available and a meaning sensible to the interpreter" [60]. For a clinician, model explainability may involve associating the learned weights of the feature space, and their relationships with outcome variables. For example, in my paper on predicting first episode BD in patients, we examined the weights given to each feature of a linear support vector machine (SVM) model. These weights represent the importance of each feature with respect to the outcome. These features included metrics of cognitive assessments that tested visual processing, working memory, executive function and language [37]. The learning algorithm

incrementally adjusted the weights of each feature to optimize a solution to detect which individuals showed cognitive abnormalities related to early stage BD compared to a control group. Using a simple, linear model, allowed us to look inside to see which features were more important for such a performance task. In this case, a clinician could conceivably interpret what the model deemed "important", but this may be only one type of explainability. However, with more data, higher dimensions, and more complicated algorithms (e.g. deep neural networks), these insights quickly become unobtainable all together [62]. This is known as the "black-box" conundrum, and it refers to human inability to rationalize how a system comes to a decision [63]. Some deep neural networks, governed by thousands of neurons, activation functions, and many hidden layers, may combine various features in such a non-linear way that no human can interpret how they might be used to reach a final prediction. This may hinder or prevent humans from trusting the output of a model. If you cannot understand why or how an algorithm comes to a decision, how can you trust it? Though studies have shown that humans trust AI as much as human domain experts, that quickly changes when an erroneous error is made [64]. Humans will convey more trust in a human's decision compared to a machine, even if the human makes more errors than the machine [64]. Even when resources are at stake, humans will risk losing more by trusting a human predictor over a machine (once they come to know of the machine's errors). This is known as "algorithm aversion", and it may be a conditioned human response for machines that occasionally make an egregious error, but are more often correct than humans [65].

Contrary to algorithm aversion, overtrust of computer models and robots can also raise concerning matters. Overtrust refers to a human-robot interaction, whereby humans remove themselves from all agency and allow the robot or AI to dictate next actions in an environment [66]. For example, some studies have shown that people willingly ignore their own intuition for escaping a simulated fire emergency in a building, and trust an evacuation robot to guide them out, even when that robot performed poorly at the beginning of the task [67]. This type of overtrust can be hazardous in psychiatry. Some believe

that the use of AI in clinical decision-making could undermine patients' trust in their physicians. Patients may feel neglected or ignored if the physician appears to rely too heavily on AI. The authors suggest that physicians need to find a balance between AI and their clinical judgment to ensure optimal patient care [68]. Additionally, there is evidence that some medical specialities (radiologists, for example), tend to overtrust AI systems and rely heavily on their outputs, leading to errors and reduced diagnostic accuracy. They suggest that radiologists should be trained to understand the limitations of AI systems and to interpret their outputs in the context of clinical practice [69]. While AI has the potential to improve clinical decision-making, it is essential to avoid over-reliance on these systems. Healthcare professionals need to be aware of the limitations of AI and understand its role in the clinical decision-making process.

The widespread adoption and benefit of AI systems in the healthcare ecosystem depend on their design to foster trust and transparency. This requirement poses a major challenge to machine learning (ML) and its application in healthcare, as humans need to comprehend and interpret these models while being cautious about their level of trust in the decisions made. To overcome this barrier, one possible solution is to prioritize simple and more interpretable models that adhere to Occam's razor principle. By doing so, humans can closely examine where mistakes are made, and reduce the impact and scope of what ML can do in psychiatry. Although this approach may limit the capabilities of ML in psychiatry, deploying complicated models in a clinical setting is futile if they are not trusted.

### 2.1.4 Conclusion

This conceptual review encapsulates the major barriers that exist in precision psychiatry today, but the problems proliferate into other fields as well. The rapid advancement and deployment of this technology has caused ripple effects which are now being noticed. The pace at which technology has moved has outpaced human understanding and planning, and these aforementioned barriers are the consequences of a rapidly changing world. Even from an evolutionary

perspective, expeditious change can disrupt the stability and equanimity of our species [70]. It should not come as a surprise that we are not fully prepared to integrate human decision-making with machines today, especially when it comes to our physical and mental health. Though precision medicine is starting to grow and evolve, it is not yet fully realized, and further considerations need to be made before we approach an optimal system. However, as we progress, solutions to these major barriers will allow for a more harmonious and trusting process with ML and technology all together.

We encourage readers to consider this conceptual review as many of these barriers will appear, in various forms, throughout the thesis. From data quality and representativeness to interpretable models, these projects reflect both the potential of ML in psychiatry, and the current barriers that exist. Some of these barriers prevented me from completing other projects. For example, we anticipated my main thesis project would be a collaboration with the Canadian military and IBM, working on a computational tool that could detect the presence of PTSD in military personnel. However, due to the difficult circumstances of COVID-19, combined with issues of data privacy and representativeness, it was not possible to complete this project. Part of my PhD journey involved facing these issues, and working with others to facilitate solutions around the barriers mentioned. Though, sometimes we were not able to overcome these obstacles, it was worthwhile to understand the ramifications of these obstacles, and how we can work to resolve them as we progress into the future. Though this was my personal account of major barriers that existed in my work, several other problems exist in the world of precision psychiatry. It is important that others working in this field consider the obstacles that face their work as well, and document them, so that we can work to understand and make room for a better future in mental health.

# Chapter 3

# Individualized identification of first-episode bipolar disorder using machine learning and cognitive tests

## 3.1 Introduction

Bipolar disorder (BD) is characterized by sporadic and recurrent episodes of mania and depression. BD has a lifetime prevalence of 2.1% worldwide [71] and according to the World Health Organization (WHO), it is one of the top 10 disorders in disability-adjusted life years (DALY) in young adults [72], due partly to its associated cognitive impairments [73]–[75]. Different meta-analyses have shown that many BD patients have neurocognitive dysfunction [76]–[78], and more severe cognitive deficits are associated with increased numbers of manic episodes and re-hospitalizations [79]. Some of these deficits can be seen in the early stages of BD, including the premorbid stage [74], however they tend to be subtle [80]. People living with BD experience a deterioration in their memory, planning, attention and learning, which negatively affects their everyday functioning. As a result, BD becomes a heavy financial and personal burden for patients, caregivers, and family members. Understanding and identifying cognitive deficits in early stage BD could eventually help clinicians intervene and prescribe accurate medication associated with the diagnostic label. This could dramatically reduce the burden of BD.

Many studies have examined the degree of cognitive impairment in BD patients in all phases of the disorder, at the "class level". During the mood episodes and euthymic phases of BD, deficits in executive control, sustained attention, verbal learning and working memory can be distinguished compared to healthy controls (HC) [77], [81]–[83]. Furthermore, several studies have indicated a positive association between the number of mood episodes and neurocognitive decline [78], [84]–[86]. The progression of BD may cause cognitive impairments to worsen, even during remission [87]. Moreover, early stage or first-episode BD (FE-BD) patients may not reveal cognitive deterioration, making early detection and intervention a challenging task [88]. This is also contingent on how long the person has been suffering from depressive symptoms before experiencing a manic or psychotic episode. This causes instability in the diagnostic process for FE patients, as some patients may be wrongly labeled as depressive prior to their first manic episode. As a result, accurate diagnoses are delayed for an average of 7.5 years after FE onset [89], [90]. Monitoring cognitive function during FE-BD is an important endeavor for improving a prognostic outlook. Identifying neurocognitive abnormalities that may be present before the onset of multiple episodes can allow the psychiatrist to treat those impairments sooner before they worsen. Available evidence suggests that patients respond more strongly to lithium if treated in the early stages of BD or before multiple episodes take place [91]. Secondly, it can inform psychiatrists about early intervention approaches [84]. Recent research has focused on training computational models using neurocognitive tests to distinguish various psychiatric disorders (including BD from MDD or schizophrenia) [92]. However, few have applied these models to the early stages of certain mood disorders, especially BD. Before such computational models can be deployed, they must be evaluated on detecting early cognitive deficits within BD alone. If successful, such tools may be used to detect if certain cognitive deficits serve as early indicators for going on to develop BD-I.

Previous studies have focused largely on group-level neurocognitive differences between chronic/multi-episode BD patients and HC. Deficits between HC and FE patients have also been examined to a lesser extent, but findings

have been inconsistent compared to chronic BD (CHR-BD). As stated earlier, some studies propose that FE-BD patients show little or no cognitive deficits compared to HC [80], [88], [93]. However, a meta-analysis done by Daglas *et al.* (2015) found mainly inconclusive results apart from one cognitive domain, working memory. Specifically, these studies consistently found a medium size effect in spatial working memory for FE patients in remission compared to HC [84]. Another meta-analysis reported by Lee *et al.* (2014) found varying size effects on processing speed, attention, verbal memory and executive function. However, their meta-analysis included all phases of BD, and the study was limited to adult participants only [94]. Together, these studies showed mixed results for group level differences between FE patients and HC. Apparently, it is difficult to detect cognitive deficits in FE patients using group level statistics. So far, we have been unable to identify any reports of individual level identifications of FE patients using machine learning techniques.

In this study, we aimed to develop a machine learning model that can identify FE patients based on cognitive tests. We trained our model on cognitive scores from a dataset of HC and CHR-BD patients, using the Cambridge Neuropsychological Test Automated Battery (CANTAB), which is a validated and standardized tool for assessment of cognitive functioning [95], [96]. We chose to train a model with CHR-BD patients, because we were interested in discovering whether ML can be used to make general identifications of BD during specific phases of the disorder. Using information from cognitive profiles of CHR-BD patients and applying them to FE-BD patients may help us discern whether cognitive deficits exist in the early stages of BD. This may help us identify cognitive markers that could be used for early detection. Machine learning models have been successfully used to distinguish later course, euthymic BD from HC with relatively high accuracy, precision and sensitivity based on cognitive scores [86], but none have applied such models on FE patients. We hypothesize that FE patients will display certain cognitive deficits that will be identified by a model trained on CHR-BD patients and HC. If our model identifies FE-BD as CHR-BD that may indicate that subtle neurocognitive deficits exist in the early phases of BD. Early identification of cognitive deficits

in FE patients could be the first step in offering a better prognostic outlook and treatment intervention, which may eventually significantly reduce the financial, emotional and medical burdens that come with BD.

## 3.2 Methods

### 3.2.1 Participants

We recruited a total of 114 patients with type-I BD from the Affiliated Brain Hospital of Guangzhou Medical University (Guangzhou HuiAi Hospital) in China. The Structured Clinical Interview (SCID) for the Diagnostic Statistical Manual (DSM-IV) was used to confirm the diagnosis of patients by trained psychiatrists, while the Hamilton Rating Scale for Depression (HAM-D) and Young Mania Rating Scale (YMRS) were performed to assess the mood state of the patients. BD patients were all right-handed and exclusion criteria included any other axis I psychiatric conditions, severe neurological or somatic illnesses, or a history of unconsciousness caused by head trauma. BD patients with a first episode manic or mixed episode was defined as the first-episode BD (FE-BD), and multiple-episode mania patients (CHR-BD) all underwent at least two manic or mixed episodes [88].

For the HC groups in each respective cohort, participants without any psychiatric illness during their lifetime were recruited from the local community by posting advertisements on the local posters. We used the SCID for DSM-IV-TR and the Axis I Disorders-Nonpatient Edition (SCID-NP) to screen them. Exclusion criteria for HC included current severe somatic or neurological illnesses, family history of psychiatric illnesses, a history of unconsciousness, any psychiatric medication currently, left-handedness, or any contraindications of MRI. General demographic information (such as sex, age and years of education) were recorded from all participants, and we also obtained clinical information like age of onset, duration of illness, subtypes, medication use from all BD participants specially. These demographic statistics can be seen in Table 3.1.

Our study was carried out in accordance with the Declaration of Helsinki.

We obtained ethics and administrative approval for our study from the Institutional Review Boards at Affiliated Brain Hospital of Guangzhou Medical University (Guangzhou HuiAi Hospital). Each participant completed the signature of written informed consent.

### 3.2.2 Cognitive Assessment

Cognitive function was assessed by using the Cambridge Neuropsychological Test Automated Battery (CANTAB) (Cambridge Cognition Limited, 2004). The tests examined domains of visual processing and memory, spatial memory, attention and executive function. Cognitive tests were administered shortly after diagnosis for those with FE-BD or CHR-BD (approximately 3 days later). The following six tests were included in our data: Delayed Matching to Sample (DMS), Rapid visual information processing (RVP), Intra-extra dimensional shift (IED), Stockings of Cambridge (SOC), Spatial Working Memory (SWM), Pattern Recognition Memory (PRM). RVP is a measure of sustained attention. PRM and DMS assess visual memory. IED, SOC, and SWM reflect executive function. Each task contains some outcome measures, and the detailed description is shown in the supplementary materials (Table 3.2).

### 3.2.3 Data Cleaning

From the original dataset with 170 variables, we dropped the following columns: IQ, season of birth, onset age, Body Mass Index (BMI), HAM-D, YMRS, suicide attempt, duration (since first diagnosis), education level and consumption of Anxiolytics, Antidepressants, Anticonvulsants, Lithium and Antipsychotics. These features were used in our demographic and clinical group analysis, but not in our machine learning analysis. This was done to ensure that no confounding variables would conflate our model evaluation. For example, the number of suicide attempts may easily explain the differences between control, CHR-BD and FE-BD participants, rendering our neurocognitive tests less vital to the explanation of our model. Additionally, we focused on only using neurocognitive

Table 3.1: **Clinical and demographic descriptions of two cohorts** [A: Chronic bipolar disorder (CHR-BD); B: First-episode bipolar disorder (FE-BD)]. One-way ANOVA tests were conducted for the variables that were continuous in nature, and chi-squared tests were used for categorical variables. Statistical significance was set at p <0.05.

| | CHR (n = 74) | | HC (n = 53) | | | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | F/Chi-square | p-value |
| Sex (Female/Male) | 44 / 30 | – | 31 / 22 | – | 2(1,127) = 0.50 | n/s |
| Age | 25.46 | +/- 0.88 | 24.09 | +/- 4.83 | F(1,125) = 1.88 | n/s |
| Onset age | 19.41 | +/- 0.73 | – | – | – | – |
| Duration | 73.26 | +/- 68.97 | – | – | – | – |
| Education (# of years) | 12.97 | +/- 2.82 | 15.05 | +/- 1.70 | F(1,125) = 16.31 | p <0.0001 |
| BMI | 22.72 | +/- 3.51 | 20.67 | +/- 2.73 | F(1,125) = 1.87 | n/s |
| HAMD | 1.70 | +/- 2.34 | 0.28 | +/- 0.77 | F(1,125) = 34.06 | p <0.0001 |
| YOUNG | 2.26 | +/- 5.66 | 0.11 | +/- 0.51 | F(1,125) = 22.75 | p <0.0001 |
| History of Suicidal Behavior | 23.00% | – | 0.00% | – | – | – |
| Psychotic symptoms | 28.30% | – | 0.00% | – | – | – |
| Anxiolytics | 10.80% | – | 0.00% | – | – | – |
| Antidepressants | 16.20% | – | 0.00% | – | – | – |
| Antipsychotics | 62.67% | – | 0.00% | – | – | – |
| Anticonvulsants | 38.67% | – | 0.00% | – | – | – |
| Lithium | 34.67% | – | 0.00% | – | – | – |

| | FE (n = 37) | | HC (n = 18) | | | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | F/Chi-square | p-value |
| Sex (Female/Male) | 18 / 19 | – | 9 / 9 | – | 2(1,53) = 0.004 | n/s |
| Age | 25.19 | +/- 7.39 | 25.00 | +/- 5.27 | F(1,53) = 0.97 | n/s |
| Onset age | 21.02 | +/- 7.07 | – | – | – | – |
| Duration | 48.04 | +/- 42.97 | – | – | – | – |
| Education (# of years) | 12.38 | +/- 3.55 | 15.27 | +/- 2.02 | F(1,53) = 3.21 | p <0.01 |
| BMI | 22.12 | +/- 3.17 | 28.99 | +/- 1.52 | F(1,53) = 1.43 | n/s |
| HAMD | 2.32 | +/- 3.56 | 0.33 | +/- 0.77 | F(1,53) = 2.34 | p <0.05 |
| YOUNG | 2.78 | +/- 5.41 | 0.22 | +/- 0.94 | F(1,53) = 2.01 | p <0.05 |
| History of Suicidal Behavior | 24.00% | – | 0.00% | – | – | – |
| Psychotic symptoms | 29.00% | – | 0.00% | – | – | – |
| Anxiolytics | 10.67% | – | 0.00% | – | – | – |
| Antidepressants | 10.81% | – | 0.00% | – | – | – |
| Antipsychotics | 73.00% | – | 0.00% | – | – | – |
| Anticonvulsants | 56.75% | – | 0.00% | – | – | – |
| Lithium | 56.75% | – | 0.00% | – | – | – |

data to better understand its predictive power in the context of identifying FE-BD and CHR-BD participants from a control group.

We also dropped data columns that were deterministically dependent on a combination of the others, such as IED total errors adjusted, IED completed stage errors, IED total trials, RVP total misses, RVP total false alarms and SWM total errors. For example, we retained the total number of false alarms in the RVP task instead of keeping false alarms from each block of the task. We further removed the variables that had over 40% missing values in each column. We also removed one individual patient who had over 60% missing variables. The remaining missing values (14) were imputed using the median of the corresponding variable, which were included in our machine learning analysis. Cohort #1 was passed to the model had 74 CHR-BD and 53 HC individuals. Cohort #2 (test set) contained 37 FE-BD (18 F, 19 M, mean age = 25.19) and 18 age- and sex-matched samples (9 F, 9 M, mean age = 24.83). Both datasets had the same 129 variables. Both datasets were standardized (using Max-Min-Scalar from the package "sci-kit learn") by scaling and translating each feature with value individually to be between zero and one.

### 3.2.4 Machine Learning Analysis

Once both sets have been cleaned and standardized, we used the data from Cohort #1 to train a linear support vector model to classify CHR-BD from HC individuals. Our ML pipeline contained a nested cross validation approach, where we used leave-one-out cross-validation (LOOCV) in the external process and 5-fold cross validation in our internal process to select the best hyperparameters based on the average accuracy of the 5 folds. LOOCV is a reliable and preferred method for datasets with a small sample size (Molinaro et al., 2005). We tuned hyperparameters based on the type of regularization and the strength of the regularization. With respect to the slight class imbalance in Cohort 1, we applied a balanced class weight by assigning a greater weight to the minority class during the internal and external cross-validation procedures. Thus, the classifier becomes more aware of the imbalanced class and adjusts the cost function accordingly. This was done to avoid generating synthetic

data. Once the hyperparameters were selected for each iteration, we estimated the generalization of the error by using LOOCV and using balanced accuracy, AUC, sensitivity and specificity as outcome metrics.

Once completed, we then fit our final model to the full data from Cohort #1. Using the same preprocessing procedure for Cohort #1, we then applied the learned model to Cohort #2 which served as our test set. The same class weight balancing used in Cohort #1 was used in Cohort #2. Here, we generated predictions for everyone in Cohort #2 that resulted in a binary classification of BD or HC. The same outcome metrics were used as in Cohort #1.

### 3.2.5 Statistical Analysis

For statistical analysis that involved examining group level differences, we conducted independent t-tests for numerical variables for the clinical and demographic characteristics across both cohorts and FE-BD and CHR-BD. For categorical variables such as sex and drug administrations, we used a Chi-squared test of homogeneity. Both tests were performed from the "SciPy. stats" package in Python 3.7.

## 3.3 Results

### 3.3.1 Group Analysis

Table 1A and 1B summarize the clinical and demographic characteristics of both cohorts. Education, HAM-D and YMRS scores significantly different from HC in both cohorts. When specifically comparing FE-BD to CHR-BD subjects, there was a significant difference in mean onset age and duration of BD. Both HAM-D and YMRS scores were not statistically different across CHR- and FE-BD patients.

### 3.3.2 Machine Learning

In our study, we trained a model using cognitive test data from 74 CHR-BD and 53 HC participants. We then applied that trained model to a Cohort #2 which included 37 FE-BD and 18 HC participants. Our learned model

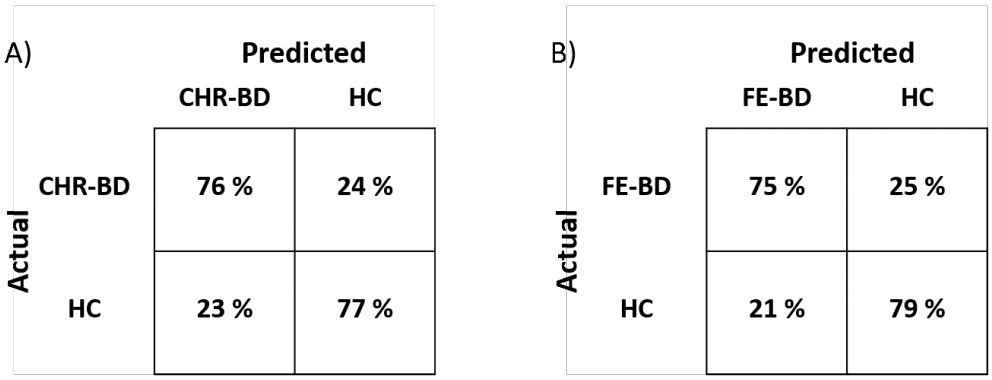



Figure 3.1: **Confusion matrix of model results for the 2 cohort datasets.** A) For the CHR-BD cohort, the precision was 82%, the sensitivity was 76%, the specificity is 77% and the overall accuracy was 77% B) For the FE-BD cohort, the precision is 87%, the sensitivity is 75%, the specificity is 79% and the overall accuracy is 76%.

on CHR-BD and HC participants scored a balanced accuracy of 77% (the arithmetic mean of sensitivity and specificity) using nested cross validation. The model selected was a linear SVC with an L1 regularization. We used that trained model to apply to Cohort #2, and we distinguished FE-BD from HC with 76% balanced accuracy, as seen in Figure 3.1. Our internal cross-validation determined that the linear SVM with a L1 regularization penalty was selected. Figure 3.1B shows a confusion matrix for FE-BD and HC. The precision score, which is the percentage of relevant instances among retrieved instances (true positives / [true positives + false positives]) was 87%, while the sensitivity was 75% and the specificity was 79%. Figure 3.3 illustrates the receiver operating characteristic curve (ROC) for the testing set (Cohort #2) and the Area Under the Curve (AUC) is 0.77. To ensure these results were not driven by a certain mood state, we observed the mood state labels retrospectively for every individual prediction in a post-hoc analysis. This was done for both cohorts. Cohort #2 showed 81% accuracy for FE-BD patients with only manic episodes (n = 26) and 54% for patients with mixed episodes (n = 11), as seen in Figure 3.4. In Cohort #1, accuracies were 72% for depressed (n = 25), 82% for manic (n = 28), 83% for hypomanic (n = 6), and 66% for

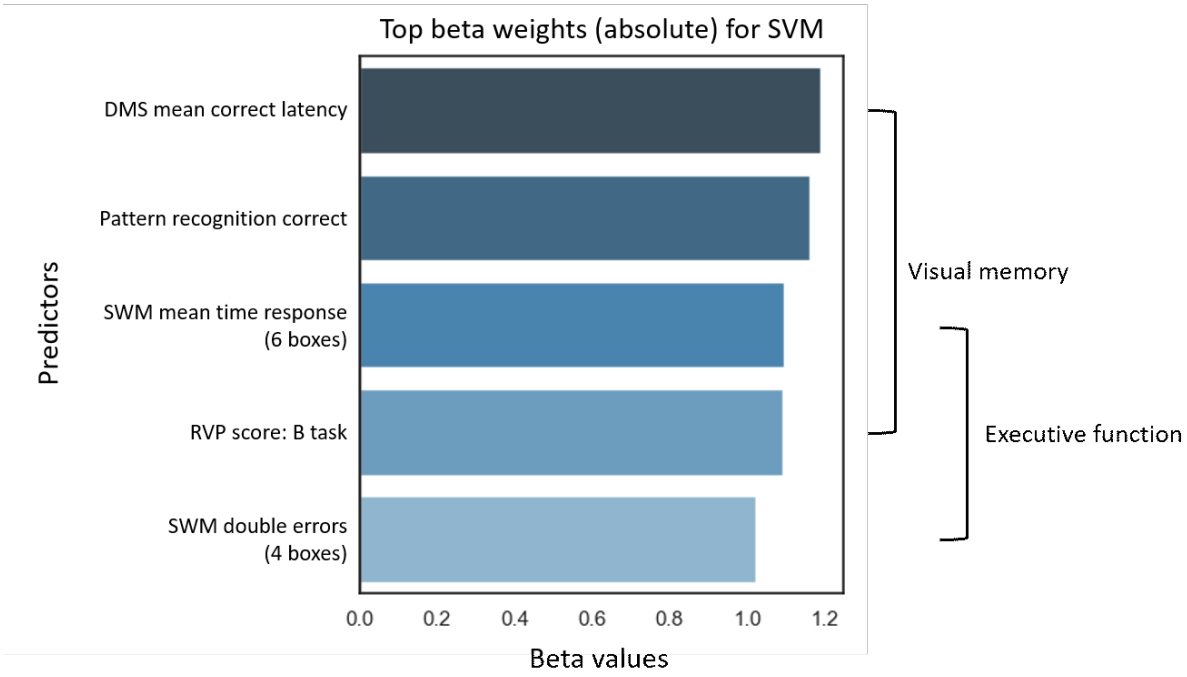



Figure 3.2: **Top predictors from Linear SVM.** A list of the 5 highest beta weights for our trained SVM model (CHR-BD vs. HC). These weights are ordered by absolute values, ranging from largest to smallest beta weights.

mixed (n = 3) and 58% for non-specified mood state (n = 12) (Figure 3.5). The largest absolute beta coefficients from our predictor variables in Cohort #1 were DMS mean correct answers (latency block), the number of correct answers in the delayed PRM, SWM mean time responses, RVPB total score, and double errors in the SWM task. These feature coefficients are shown in Figure 3.2. Of the top 5 predictors, 3 are from the visual memory domain and 2 from executive function. A full list of predictors and their beta weights are shown in Table 3.3.

## 3.4 Discussion

In this study, we built a model that could identify FE patients based on different dataset involving CHR-BD patients. These cognitive tests examined visual attention, memory and executive function. We trained the model to make

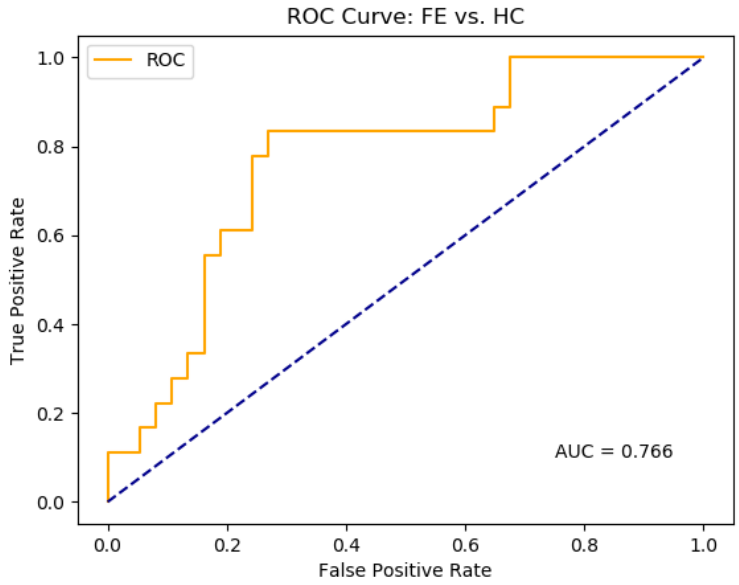



Figure 3.3: **Receiver operating characteristic curve (ROC) results from testing set.**

individualized identifications of FE patients and HC. We examined whether cognitive test scores in FE patients look like CHR-BD patients. Our linear SVM identified individual CHR-BD patients from HC with 77% balanced accuracy. When applied to Cohort #2, our model achieved a balanced accuracy of 76% (AUC = 0.77), a sensitivity of 75%, and a precision score of 87% which illustrates that our model is generalizable to the various stages of BD. These results were also fairly consistent across depressed, manic and hypo-manic mood states in the Cohort #1, and higher for manic mood states Cohort #2 [97]. These results suggest that early cognitive markers can be individually identified in FE patients, even if these deficits are subtle or comparable to a control group [80], [88], [93].

Making individualized predictions using neurocognitive tests on FE patients yields many psychiatric and clinical possibilities. Firstly, while previous studies have examined which factors are most predictive of a first manic onset, our study used a trained model of CHR-BD patients to identify FE patients and

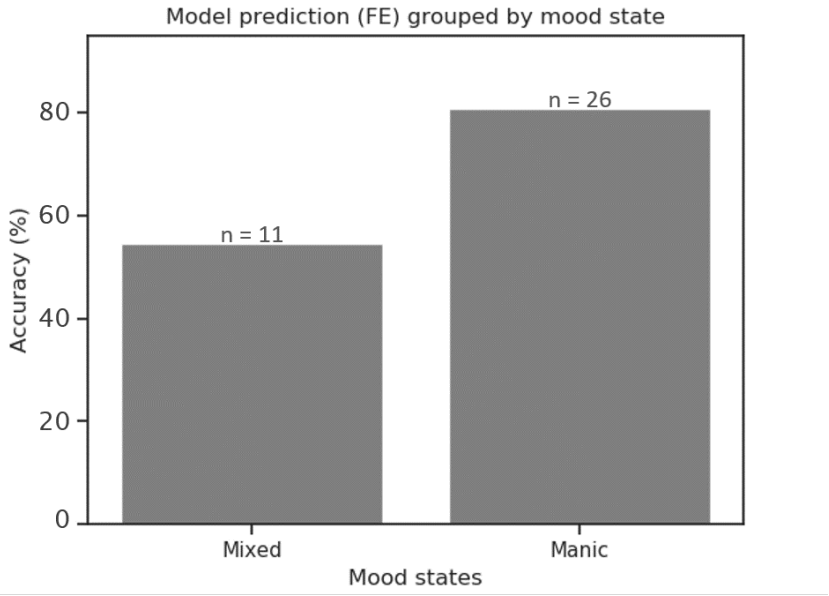



Figure 3.4: **Model prediction (FE) grouped by mood state.** A summary of the test set results (FE-BD) grouped by mixed, manic episodes or HC. The x-axis represents the two sub-categories of FE-BD patients, and the y-axis is the classification accuracy within those two groups.

which factors may be most predictive. This shows that persistent neurocognitive deficits may be used to predict the presence of early manic episode. Detection of FE patients using several neurocognitive tests may yield value for understanding the subtle cognitive deficits that may otherwise not be detected from a control group. Tracking the progression of cognitive functioning in BD could also assist clinicians in providing accurate, individualized treatments. Secondly, our data consists of cognitive tests that are accessible, versatile and inexpensive form of data compared to conventional techniques such as fMRI, sMRI or Diffusion Tensor Imaging (DTI). With the accessibility of this data, there may be an opportunity to collect more data at different time-points during BD. This could be beneficial for monitoring cognitive dysfunction, allowing for psychiatrists to gauge cognitive symptoms over the course of BD. From an ML perspective, collecting more data can strengthen the confidence and predictions of a model. Thirdly, cognitive assessments have recently gained traction as a

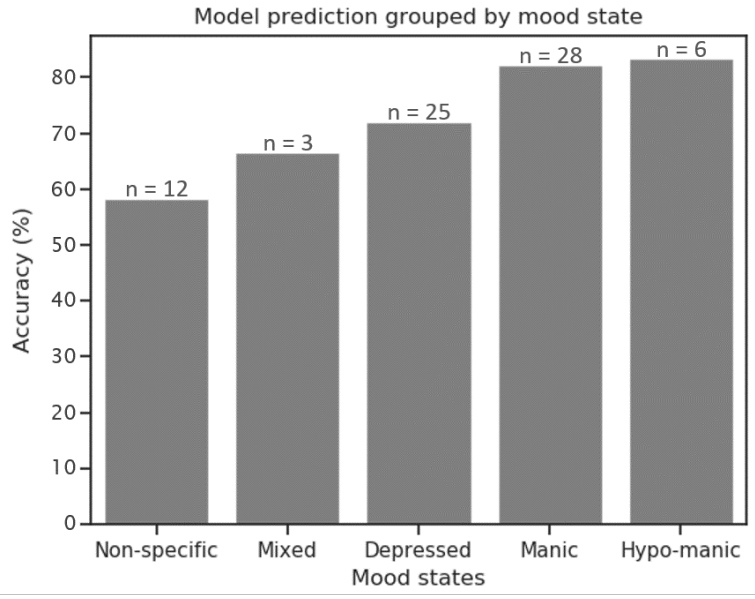



Figure 3.5: **Model prediction (CHR-BD vs. HC) grouped by mood state.** A summary of the training set accuracies (CHR-BD vs. HC) grouped by the categories of mood state. The x-axis represents the sub-categories of CHR-BD patients, and the y-axis is the classification accuracy within those two groups. Accuracies were highest for those who were suffering from hypomanic and manic episodes.

critical predictor of functional outcomes in disorders such as major depressive disorder and BD [98]. In some studies, cognitive measurements have been used to objectively identify individuals in need of therapeutic intervention while also serving as an indicator for various stages of psychiatric disorders including early and remission phases [98], [99]. In BD, cognitive dysfunction remains one of the largest factors affecting quality of life [100], [101]. Thus, cognitive testing should be considered along the side of clinical assessments to further improve treatment of cognitive deficits as well as an indicator of functional outcomes.

It is also important to consider the mood states of the patients at the time of cognitive tests, as those mood states can be a major factor in discerning cognitive impairments of both CHR-BD and FE [79]. Cognitive function seems

to be impaired during acute mood phases of BD. For most cognitive measures, studies comparing neuropsychological functioning across different mood states [79], [102] found no significant impairments between different mood states. In our study, we compared our model performance across mood states. Figure 3.4 and 3.5 in the supplementary materials reveal accuracies of patients within a certain mood state, showing that, for both of cohorts, accuracies for depressed, manic, hypomanic were largely unaffected. However, in both cohorts, mixed mood-states reflected lower accuracies, which is contrary to other studies [79], [102]. Sample sizes for mixed mood-states were very low compared to others (n = 12 for FE-BD, n = 3 for CHR-BD) which may have affected statistical power. Yet, accuracies in both cohorts were not significantly different from one another.

We found that the performance in visual memory tasks and sustained attention could potentially be predictive to mania. This suggests that visual memory and attention are affected by people suffering from type-I BD. More importantly, these predictors were used to identify FE patients at high accuracies, suggesting that these patients are more prone to errors in visual memory and attention tasks. Findings on cognitive impairments in BD have been well demonstrated by many cross-sectional studies [103]–[105]. Generally, BD patients exhibit more visual memory and attention deficits compared to HC [103], and these abnormalities do not seem to be state-dependent. These deficits can be seen in all phases of BD [82], [104]. Beyond that, there is also evidence that a more extensive history of psychosis in BD and schizophrenic (SCZ) patients is correlated with working and visual memory impairments. Frydecka *et al.* 2014 compared cognitive impairments between BD and SCZ patients using visual and working memory tasks like DMS. With respect to between group differences, they observed no differences between SCZ and BD groups, but found a strong correlation between cognitive performance and recorded history of psychotic symptoms. This finding reinforces the idea that cognitive dysfunction is less pronounced in the early stages of BD, making it difficult for psychiatrists to become aware of any deficits. However, according to our results, they are not absent and can be detected using ML techniques that

are more sophisticated than conventional statistics. Another study by Glahn *et al.* 2007 compared BD patients with a history (BD-H) of psychotic symptoms to BD patients without such history. BD-H patients showed greater impairment to executive function and spatial memory tasks compared to non-psychotic BD patients at a class level. Some of these impairments may exist due to a history of manic/psychotic episodes. Here, it seems that CHR-BD and FE patients both have impairments in visual and working memory, and our model was sensitive to these impairments.

In terms of attention and psycho-motor speed, our model relied on RVP scores to identify FE patients. There is evidence that BD affects visual information processing systems [107]–[109]. One study examined visual processing using three different measures in BD, SCZ and HC [107]. They examined rapid visual processing at an early stage (a 0-100ms backwards masking), a middle stage (100-200ms object substitution), and a late stage task (200-500ms RVP). In the early and middle stages, they found that SCZ patients displayed deficits in all 3 stages of visual processing, but BD patients only showed disruptions in the later stage RVP task. This suggests top-down, higher-level cognitive functions are more pronounced in BD patients compared to SCZ patients, who display more low-level, early visual processing deficits. Early, middle and late stage metrics for RVP tasks may be necessary to help identify and assess the severity of visual processing impairment. Perhaps more manic episodes affect early, low-level visual processing whereas fewer manic episodes inhibit top-down, later onset visual processing. An interesting future study could involve FE-BD and CHR-BD patients performing an RVP task where early, middle and late stage metrics are recorded, but this was not done in our study. In addition, visual processing deficits are also correlated to history of manic symptoms as well in BD. More manic episodes result in reduced performance on visual processing tasks [110]. Nevertheless, the top predictors from our model seem to align with the literature on which specific cognitive deficits exist in CHR-BD patients. While it is difficult to ascertain whether FE patients reveal the same degree of impairment, our model was able to make the distinction that they resemble CHR-BD patients more than HC. This may be an important

distinction, since there is debate about whether any cognitive deficits exist in FE patients.

Our current study has several limitations. Firstly, our overall sample size is small. Although we validate our model prediction on a separate cohort of FE-BD patients, the outcome still needs to be interpreted with caution. Secondly, while our mood state analysis reinforces our findings that mood state may not be a determining factor of our results, the sample sizes are not large enough to make that inference. Considering the current sample size, accuracies for depressed, manic and hypo-manic mood states remained high (above 70%), but the mixed and non-specific mood states suffered in accuracy. Both cohorts showed significantly lower accuracies for mixed patients (54% for FE-BD, 66% for CHR-BD). It is noteworthy that only 11 subjects were in a mixed mood state for Cohort #2 and 3 for Cohort #1, and the chi-squared test for both cohorts revealed no significant difference between mood-state accuracies. More patients with varying mood states must be recruited to determine whether mood state is a confounding factor in predicting cognitive markers in FE patients. Lastly, it is difficult to speculate on the top coefficients from our model since there is a steady diminution of all predictors as seen in the supplementary materials. Future longitudinal study with individuals at risk of BD may help to confirm the predictive power of these potential predictors.

In summary, our study successfully distinguished FE patients from HC with a 76% accuracy using cognitive test scores from the CANTAB neurocognitive battery. This model was trained on data from CHR-BD and HC participants, which distinguished CHR-BD and HC at an accuracy of 77%. These results suggest that cognitive impairments do exist after only one manic episode for FE patients, as our model was able to identify that. Several features from our model reflected the same cognitive deficits identified by other studies. These features used by our model served to identify FE patients at an individual level. From the perspective of precision medicine, the application of ML may offer insights for psychiatrists in their attempt to diagnosis BD more accurately. This is a worthwhile endeavor because BD is largely mistaken for other psychiatric disorders [111]. Psychiatrists have to make important

decisions that can involve numerous factors such as symptomatology profile, previous efficacy, medical comorbidities, family history, cognitive deficits and so on [112]. This study provides a first step into using cognitive markers to help validate the diagnostic label of BD, especially in the early stages. With this, our hope is that future computational models will help further identify and validate cognitive differences between various psychiatric disorders such as BD and major depressive disorder. In that way, psychiatrists can make informed decisions on early identification of BD, which may lead to a better prognostic outlook.

### 3.4.1 Supplementary Tables

Table 3.2: **Cognitive outcome measures.** A glossary of terms from the CANTAB manual. Each cognitive domain is broken down into the separate tests and each test has a list of variables that were recorded.

| Function | Task | Index |
|---|---|---|
| Visual Memory | DMS | DMS Prob error given correct |
| | | DMS Prob error given error |
| | | DMS A' |
| | | DMS B" |
| | | DMS Mean correct latency |
| | | DMS Mean correct latency (all delays) |
| | | DMS Mean correct latency (simultaneous) |
| | | DMS Total correct |
| | | DMS Total correct (simultaneous) |
| | | DMS Total correct all delays (0, 4000, 12000ms delays) |
| | | DMS Percent correct |
| | | DMS Percent correct all delays |
| | | DMS Percent correct simultaneous (0, 4000, 12000ms delays) |
| | PRM | PRM Mean correct latency [immediate] |
| | | PRM Number correct [immediate] |
| | | PRM Percent correct [immediate] |
| | | PRM Mean correct latency [delayed] |
| | | PRM Number correct [delayed] |
| | | PRM Percent correct [delayed] |
| Sustained Attention | RVP | RVPA (Blocks 5-7) |
| | | RVPB (Blocks 5-7) |
| | | RVP Total hits (Blocks 5-7) |
| | | RVP Total misses |
| | | RVP Total false alarms (Blocks 5-7) |
| | | RVP Total correct rejections (Blocks 5-7) |
| | | RVP Probability of hit (Blocks 5-7) |
| | | RVP Probability of false alarm (Blocks 5-7) |
| | | RVP Mean latency (Blocks 5-7) |
| Executive Function | SOC | SOC Problems solved in minimum moves (2-5 moves) |
| | | SOC Mean moves (2-5 moves) |
| | | SOC Mean initial thinking time (2-5 moves) |
| | | SOC Mean subsequent thinking time (2-5 moves) |
| | SWM | SWM Between errors (4,6,8 boxes) |
| | | SWM Within errors (4,6,8 boxes) |
| | | SWM Double errors (4,6,8 boxes) |
| | | SWM Total errors (4,6,8 boxes) |
| | | SWM Strategy |
| | | SWM Mean time to first response (4,6,8 boxes) |
| | | SWM Mean time to last response (4,6,8 boxes) |
| | | SWM Mean token-search preparation time (4,6,8 boxes) |
| | IED | IED Pre-ED errors |
| | | IED EDS errors |
| | | IED Stages completed |
| | | IED Total errors |
| | | IED Total errors adjusted |
| | | IED Completed stage errors |
| | | IED Errors (block 1-9) |
| | | IED Total trials |
| | | IED Total trials (adjusted) |
| | | IED Completed stage trials |

Table 3.3: **List of model beta weights from linear SVM.** Each predictor is associated with an outcome variable from the cognitive tasks listed in Table S1.

| Model predictors | Beta weights |
|---|---|
| PRM Number correct delayed | 1.162394234 |
| RVPB | 1.090322917 |
| SWM Double errors 4 boxes | 1.021126761 |
| DMS Total correct 12000 ms delay | 0.916837052 |
| SWM Within errors 4 boxes | 0.813712671 |
| Prmpct | 0.734294903 |
| SOC Problems solved in minimum moves (2 moves) | 0.673512728 |
| PRM Number correct immediate | 0.611912419 |
| IED PreED errors | 0.583026043 |
| soc3 | 0.572376868 |
| SOC Mean moves (3 moves) | 0.572376868 |
| DMS clad | 0.521811914 |
| 11 DMS Mean correct latency all delays | 0.521811914 |
| IED Errors block1 | 0.46466718 |
| IED Errors block3 | 0.437218341 |
| RVP Total correct rejections block 7 | 0.33614977 |
| SOC Problems solved in minimum moves (3 moves) | 0.330945598 |
| RVP Total correct rejections block 6 | 0.327045315 |
| IED Errors block 5 | 0.325601954 |
| IED Errors block 2 | 0.320657017 |
| RVP Probability of hit block 7 | 0.290468019 |
| RVPA block 7 | 0.287379752 |
| SOC Mean subsequent thinking time 3 moves | 0.267308373 |
| RVP Total hits block 7 | 0.229759097 |
| DMStcal | 0.202059851 |
| DMS Total correct all delays | 0.202059851 |
| DMS percent correct all delays | 0.202059851 |
| RVP Mean latency block 5 | 0.17047578 |
| IED Errors block 7 | 0.165524851 |
| RVP Probability of hit block 6 | 0.156551379 |
| RVP Total hits block 6 | 0.156551378 |
| RVPB block 7 | 0.137209424 |
| RVPB block 6 | 0.135684752 |
| SOC problems solved in minimum moves | 0.134332101 |
| RVP Total correct rejections block 5 | 0.13265144 |
| SWM Within errors 6boxes | 0.122193167 |
| RVPA | 0.10819042 |

# Chapter 4

# Predicting pediatric anxiety from the temporal pole using neural responses to emotional faces

## 4.1 Introduction

Clinical anxiety is associated with inability to control or auto-regulate one's autonomic response [113], and is the most common mental illness among children and young adults [114], with a lifetime prevalence rate of 28.8%[115], [116]. The median age of onset for all anxiety disorders, at 11 years old, marks this as the earliest among all psychiatric disorders and over 30% of pediatric cases meet criteria for two or more subtypes [114], [116]. Despite high prevalence and possible early onset, these disorders are often under-reported because of conflation of normal developmental-behavioral patterns with anxiety symptoms. Assessment is typically limited to diagnostic interviews and questionnaires to produce a diagnostic label, which comes with its own validity issues [115], [117], [118]. Anxiety and related symptoms may have profound effects on neurological functioning in a child's rapidly developing brain [117], [119] and, over extended periods of time, may lead to cognitive, social, and emotional deficits [113]. For example, adolescents with high trait anxiety exhibit an attentional bias (*i.e.* pay greater attention) to negatively valenced faces [120], [121]. Although socioemotional circuits in the brain have

been implicated in numerous psychiatric disorders, including anxiety [122], such cognitive deficits have rarely been used as an indication of brain mechanisms underlying psychopathology.

Cognitive models of anxiety suggest that negative biases exist for performance on information-processing tasks [123] — in particular, anxious individuals allocate greater attention to negative or threatening stimuli [121]. They may find threatening words more salient, and may remember them more often than non-threatening words [124], [125]. Emotional facial expressions are often perceived as more negative or threatening (even if they are typically judged as neutral), and this is associated with activation of affective brain circuits [120], [126]. These attentional and perceptual biases are thought to be an important feature underlying the etiology of anxiety disorders, a view supported by functional neuroimaging studies. Despite the likely clinical significance of these biases, few studies have focused on the adolescent population during a facial emotional processing task [127]–[129], and rarely have machine learning methods been applied to assess if neural signatures underlying such biases may be used to identify children suffering from anxiety.

Functional neuroimaging measurements during facial processing tasks have helped reveal neurological underpinnings of emotional regulation. Overall, there is evidence of dysregulated fear-circuitry related regions, including the amygdala and prefrontal cortex (PFC) [130]. Children with panic disorder (PD) or generalized anxiety disorder (GAD) may exhibit exaggerated amygdala responses to fearful faces compared to non-anxious or depressed children [131]. Hyperactivity has been observed in several limbic brain regions in separation anxiety disorder (SAD) patients when responding to fearful faces, including the fusiform gyrus (associated with facial recognition), and there is evidence of increased connectivity between the fusiform gyrus and amygdala, as well as the fusiform gyrus and superior temporal sulcus [132]. Also, abnormal neural responses to emotional faces have been reported for adults with GAD, PD and SAD, with greater right amygdala activation reported in response to fearful versus happy faces[133]. From these studies, similar amygdala activation patterns to happy faces were reported for both patients and controls, indicat-

ing that this area is also responsive to positively valenced facial expressions. Increased responses in the superior temporal sulcus, an important area for deriving social and emotional information, were observed for SAD and PD patients viewing fearful faces. The findings mentioned above have focused mainly on between group differences or similarities. In recent years, advanced data analysis methods, such as machine learning, have enabled accurate prediction on an individual basis [134]. This approach holds the potential to enable improvement of clinical decision-making (such as diagnostic assessments), and to provide evidence-based determination of which brain regions display the largest differences between individuals in different classes (*e.g.* diagnosis versus no-diagnosis cases), based on fMRI data, while the participants perform passive or active tasks. In this study, we explore whether machine learning analysis of data in the facial recognition paradigm, may allow us to identify, with higher precision, which children will suffer from anxiety.

Conventional neuroimaging analysis of two different populations (*i.e.* anxious versus non-anxious) involves comparing neural activation of various regions between the groups, anticipating that comparing the blood oxygenation level-dependent response (BOLD) at specific voxels will show significant differences. However, this analysis mainly focuses on univariate and group-level statistics and may not lead to predictions for individual cases due to the overlap of neural responses at any given voxel [134], [135]. Multivoxel pattern classification (MVPA) applies machine learning algorithms to fMRI BOLD signals to produce predictive models [136]. These models can categorize brain patterns into distinct stimulus conditions (*i.e.* emotional faces) or groups based on spatial and temporal discriminative neural signatures from high dimensional neuroimaging data [135]. This analysis can also reveal which brain regions differ the most between two groups or stimulus conditions. Neural signatures can be further clarified with advanced alignment techniques like the probabilistic shared response model (SRM), which aligns patterns of neural responses across subjects into a common, lower-dimensional space [137]. Here, we demonstrate that MVPA can be used to decode brain patterns related to the disease state of adolescent children. MVPA may also indicate which brain regions are key

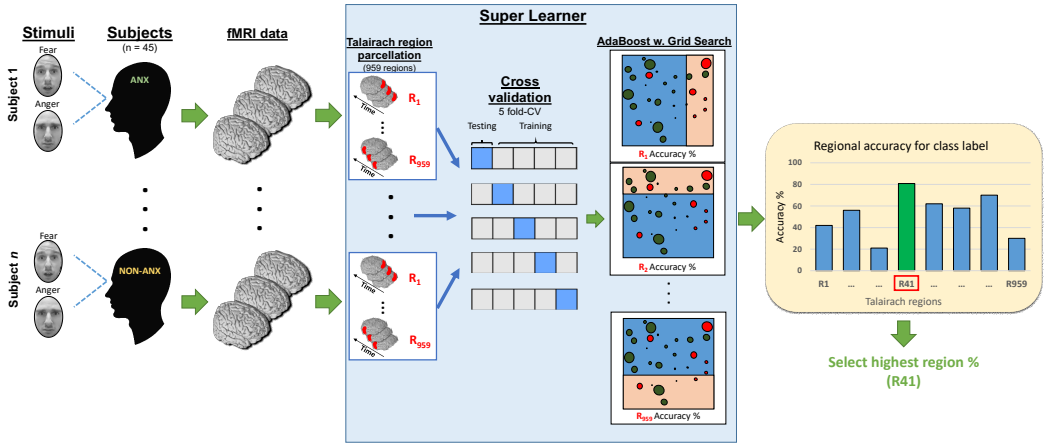



Figure 4.1: **Processing pipeline for selecting the best Talairach region.** Super learner (SL) parcellates the whole-brain data into 959 regions based on the Talairach atlas. It determines which region can best distinguish between anxious and non-anxious children. The SL uses a nested cross-validation process to hyper-tune parameters for an AdaBoost model. The SL uses the regions as a hyperparameter within this process. The region with the highest average accuracy was selected for our analysis. Note: each time point produced its own prediction; we labeled each person with the majority vote over the time points for that subject.

aspects for altered functional connectivity in anxious children in this context.

Using a publicly available dataset (`https://openneuro.org/datasets/ds000144`) consisting of task-based fMRI data from children with anxiety disorders such as SAD, SP and GAD: (1) We applied a data driven approach to determine a combination of brain regions to distinguish anxious versus non-anxious children with above chance accuracy based on facial-emotional processing. (2) We examined neural correlates of angry and fearful faces to distinguish those stimuli using similar techniques. Figures 5.2 and 4.6 illustrate the analysis and posthoc pipeline for these research questions (refer to the online methods for a full description of the study).

Our approach is based on task-based fMRI data rather than resting-state MRI [138]. A key question is whether task-based fMRI derived regions can be

linked to various resting-state networks in this context. Many reports indicate that a functional imbalance in large scale networks, such as the default mode network (DMN), the salience network, or the affective network, play a crucial role in anxiety disorders [139]–[141]. However, we found that intrinsic resting-state network activity may not differ significantly from task evoked responses, in accordance with several sources suggesting that task-based responses are related to modest changes compared to intrinsic activity [142], [143]. If certain regions arise as significant predictors of childhood anxiety using machine learning analysis for the task-based approach, it will be important to compare them with components of resting-state networks previously associated with anxiety.

Research has not yet established a clear link between brain-behavioral function and clinical diagnosis in children, which is problematic. The hope is that research into developmental psychopathology will bridge the gap between psychiatric practice and neuroscience [144]. Our current approach may enable us to relate functional brain measures to pediatric diagnoses in anxiety disorders, and may also help to generate new therapeutic insights.

To date, we are not aware of published attempts to use machine learning to validate psychiatric disorders in young children (in this case, 5-10 years old) using task-based fMRI data. We propose that distinguishable neural substrates in anxious vs. non-anxious children can be identified with our machine learning approach to task-based paradigm fMRI analysis for individual predictions on a case by case basis.

## 4.2 Methods

### 4.2.1 Data and Code

This paper analyzed data provided by Carpenter, K.L., Angold, A., Chen, N.K., Copeland, W.E., Gaur, P., Pelphrey, K., Song, A.W. and Egger, H.L. (2015), who posted their data-set on `https://openneuro.org`. The link to the data repository is `https://openneuro.org/datasets/ds000144`. Their data was made publicly available on 2018-03-26 [145], [146]. All analysis can be replicated using our GUI-based toolbox, easy fMRI. The GitLab repository

can be found and cloned at `https://easyfmri.learningbymachine.com/`.

### 4.2.2 Recruitment

Secondary analysis of existing data was obtained from Carpenter et al., (2015). Children were initially recruited from the Duke Preschool Anxiety Study (DPAS), which was a longitudinal, multi-phase study. The last phase was entitled "Learning about the Developing Brain study" (LABD), where 208 children who participated in previous phases of the DPAS were recruited to take part in this study, which examined brain development in children suffering from anxiety. Of the 208 children, 155 were eligible to participate in the neuroimaging phase. Children who met the criteria for generalized anxiety disorder, SP, and/or SAD were recruited into the "case" group, and children who did not meet the criteria for an anxiety disorder were recruited as the comparison group. Children in the LABD were not excluded for comorbid non-anxiety disorders or for taking psychotropic medications [145].

Parents completed the Preschool Age Psychiatric Assessment (PAPA) for children involved in this study [147]. The PAPA is a diagnostic instrument for assessing psychopathology of children aged 2-9, and it is based on the parent version of the Child and Adolescent Psychiatric Assessment [145], [148]. Frequency, duration, and the onset of symptoms are collected to determine whether the child meets the diagnostic criteria for anxiety disorders in the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV). The PAPA assesses symptom severity during the previous 3 months, as shorter recall periods have been shown to reflect more accurate recall [148]. A composite score of GAD, SP, SAD, and depression symptoms were obtained from the PAPA and was used as a measure of school-age emotional symptomatology [145].

This study was approved by the Duke University Medical Center Institutional Review Board, and was carried out in accordance to U.S regulatory requirements related to the protection of human research participants, which include the Accreditation of Human Research Protection Program (AAHRPP) and the Health Insurance Portability and Accountability Act (HIPAA) guide-

lines. Verbal assent from the child and informed consent from the parent were obtained after a full description of the study was presented. Children and parents were financially compensated with gifts. or money vouchers [145].

### 4.2.3 Participants

Children eligible for the fMRI study had to meet 3 requirements: (1) They completed the first phase of the DPAS study, (2) they must be older than 5 and half years old, (3) have successfully completed a mock scan session in the MRI machine. Children were placed into one of two groups. The first group involved anxious children, who met the criteria for GAD, SP SD, or some combination of the 3 using the PAPA questionnaire, and non-anxious children who served as a control group. Of the 155 children initially recruited, only 45 had usable data due to a number of reasons including parents or child refusal to take part, absentees, excessive motion in the scanner, and lower IQs [145]. Of those 45 children, 22 were in the anxious group and 23 were in the non-anxious group. The anxious group contained individuals with either one or more anxiety disorders. Within the anxious group, 15 children met the criteria for GAD, 11 for SP, and 10 for SAD. 12 out of 22 anxious children met the criteria for more than one anxiety disorder. The age range of both groups was between 5.5-9.5 years old, as seen in Table 4.1. Impairment and emotional symptoms were recorded prior to the start of the fMRI study and were representative of psychiatric symptoms that interfere with daily functioning. Impairment scores were assessed using the World Health Organization's International Classification of Functioning, Disability, and Health [149]. Emotional symptoms were measured on a composite scale that accounted for both anxiety and depressive symptoms[145].

### 4.2.4 Functional MRI task

The fMRI task was a block design, emotion face processing task. Facial stimuli from the NimStim Stimulus Set were selected (45), but only angry and fearful faces were used according to Carpenter et al., (2015). Each subject completed 2 blocks. At the beginning and end of each run, there was a 16-second fixation

block and 15-second task blocks were stationed in between and separated by 12-second baseline fixation blocks, which consisted of a colored star in the center of the screen. Faces were shown for 1.25 seconds with no inter-stimulus interval. Each run contained 3 blocks of fearful and angry faces exclusively, with the order of the emotional faces randomized. To make sure the children were staying engaged, they were instructed to press a button whenever a face with glasses was shown on screen. These faces were randomly placed throughout the blocks and expressed the same emotion as the other pictures within the block. The average task accuracy was 83.33% for non-anxious children and 82.29% for anxious children. The study used a block-design fMRI scheme in which all participants viewed the same number of stimulus presentations and consistent interstimulus intervals. This experiment scheme allowed us to extract 35 time points (based on our design matrix) for each participant during preprocessing. Further, these time points were also temporally aligned to ensure $m^{th}$ time point for all participants represented the same type of stimuli.

### 4.2.5 MRI acquisition

MRI acquisition was completed on two different 3T GE scanners. Of the 45 participants, 15 (8 anxious, 7 non-anxious) were scanned using the EXCITE HD system, and 30 participants (14 anxious, 16 non-anxious) were scanned on the MR750 system. Parameters and pulse sequences were congruent between the two systems, and calibration metrics such as spatial accuracy and dynamic signal stability were validated using an agar phantom (soft tissue mimic). In both systems, scans lasted 5 minutes and 44 seconds and 172 functional images were generated during the task. For each run, between 34-39 slices were generated which were parallel to the AC-PC plane using a BOLD-sensitive EPI sequence (voxel size: 4 $mm^3$; Repetition time: 2000ms; Echo time: 27ms; Field-of-view: 24 cm; Flip-angle: 77; Interleaved-odd acquisition) [145]. Co-registering the functional images was done in conjunction with a high resolution T1-weighted anatomical scan using the 3D-FSPGR sequences with SENSE (voxel size: 1 $mm^3$; Repetition time: 8.096ms; Echo time: 3.18ms; Inversion time: 450ms; Field-of-view: 25.6 cm; Interleaved-odd acquisition) [145]. Batch effects were

recorded as a covariate in the machine learning analysis to ensure manufacturing differences between the two systems were not a cause of functional differences.

### 4.2.6 Pre-processing

Data was pre-processed and analyzed using Easy fMRI (version 1.8B8800) (`https://easyfmri.github.io.`) and FMRIB Software Library (FSL version 6.0.3). We have used the "fMRIPrep" pipeline [150], [151] — which includes brain extraction, registration to standard space, motion correction, slice time correction, normalization, and spatial smoothing. To prepare the images for registration, we first used the Brain Extraction Tool (BET) to eliminate non-brain tissues such as the scalp and brain marrow. We then registered all the subject's brain images to a common reference coordinate system using the MNI-152, 2 mm resolution (T1 weighted) standard space. To anatomically align the brain images, we used an affine (12 degrees of freedom, 12 DOF) transformation to rotate, translate, and scale the images into alignment [152]. Motion correction was also handled in this affine transformation. Because of movement in the scanner, we needed each voxel to correspond to a consistent anatomical point for each point in time [152], [153]. Here, we chose to use the first image in the time frame to reference all other volumes at other time points. Fortunately, the dataset we acquired already removed excess motion subjects. Carpenter et al., (2015) removed relative and absolute motion and intensity jumps greater than three standard deviations from the run mean as part of their scrubbing protocol. The mean of runs was determined by taking the absolute deviation relative to the mean of runs after each voxel was passed through a high pass filter to remove low-frequency drifts (1/60 Hz) [145]. Task blocks were removed from analysis if two volumes were removed from the start of the block or more than 3 volumes in total were removed from the block. Additionally, the entire run was excluded from subsequent analyses if more than one block of emotional stimuli was removed [145]. Next was spatial smoothing. Spatial smoothing is a method used to increase the signal-to-noise ratio in fMRI brain volumes. Smoothing was done by using a 3D convolution with a Gaussian kernel to replace voxel intensities with a weighted average of

neighboring intensities. We specified our Full-Width-Half-Maximum (FWHM) kernel to be 5.0 mm. After, we applied a global intensity normalization between subjects and sessions. Lastly, we used temporal filtering, which is a removal of high and or low frequencies in the raw signal of voxel intensities via band-pass filters. In a time series of each voxel, there may be scanner related or physiological signals that cause high-frequency noise.

### 4.2.7 Analysis: Anxious versus non-anxious classification

#### Machine learning analysis

Using the Talairach atlas (with 2mm voxel size), we used a super learner (SL) that segmented brain regions (959) and used them as hyper-parameters to examine which areas could best separate our diagnostic labels. Figure 5.2 illustrates the full machine learning pipeline for our primary analysis. All machine learning analysis was done using Python 3.9[154] or in Easy fMRI [151]. Subsequent libraries included scikit-learn (version 0.23.1) [155], SciPy (version 1.6.1) [156], Pandas [157], and Numpy (version 1.20.1) [158].

A SL is seen as an "ensemble of ensembles" that combines models or model configurations on the same split of data, and then uses out-of-fold predictions to select the best configurations or models [159]. We applied the whole-brain data to a SL — where it parcellated the neural responses based on the Talairach atlas and then found an optimal prediction model for each of the regions. The SL returns the model which can best distinguish class labels (anxious versus non-anxious). The SL would make a prediction on every time point for each subject (35 time points), then use a majority vote to make a final prediction for the class label associated with that individual, regardless of task stimuli. We split the data using 5-fold CV based on the participant IDs (36 / 45 participants were considered our training set, and 9 / 45 our testing set). The SL used a nested cross-validation process, whereby the outer CV process used $4/5^{\text{ths}}$ of the participants for a training set (balanced for diagnostic labels), and $1/5^{\text{th}}$ for a testing set. Within the training set, another 5 fold-CV was used to fine tune the hyper-parameters within the SL. Folds were split based on the participant

ID instead of time points (30 / 36 participants were in the training set, and 6 / 36 were in the validation set). An ensemble classifier (AdaBoost) with a logistic regression base estimator was used to make the predictions on each participant.

AdaBoost (short for Adaptive Boosting) is an ensemble machine learning paradigm where multiple models (often called "weak learners") are amalgamated in such a way to achieve more robust results [159]. This is done by setting weights for both the weak learners and the data points. The algorithm forces the weak learners to concentrate on observations that are difficult to classify correctly. AdaBoost uses boosting, a sequential ensemble process that gives misclassified cases a heavier weight, samples without replacement, and reduces the bias-variance tradeoff by combining weak or shallow learners together in a voting process to make predictions[159]. We used four hyper-parameters to tune our classifier within the inner CV. Hyperparameter tuning was done using GridSearchCV, a function within Scikit-learn [155]. First, we used a different number of estimators [n_estimators = 10, 50, 100, 150] to determine the maximum number of estimators at which boosting is terminated. Second, we adjusted the learning rate [learning_rate = 0.05, 1, 2]. Third, we changed the number of max iterations completed by AdaBoost [max_iterations = 100, 500, 1000]. Next, we tuned the type of regularization performed by the logistic regression estimator [penalty = L1, L2, none]. This was also coupled with the regularization penalty variable [C = 0.5, 1, 2]. Lastly, we used the segmented regions from the Talariach atlas as a hyper-parameter. We evaluated the performance of our results by using the accuracy, which was computed as the average accuracy across the folds in the outer CV. We used precision, recall and F1-score to evaluate the final model.

### Statistical analysis of top region(s)

We conducted a high level, between-group analysis for the ROI selected to examine activation differences for anxious versus non-anxious children. Instead of using a regular classification analysis, we conducted a grouped Bayesian representational similarity analysis (GBRSA) that can compare the (dis)similarities

between different cognitive states across multiple participants. This was done to determine whether the pattern of activity between anxious and non-anxious children were statistically different in region #41.

RSA is a similarity fMRI analysis method that explores the neural response patterns of brain regions across different stimuli or different groups[160], [161]. Using a measure of similarity or dissimilarity (1 - measure) such as euclidean distance, Spearman's correlation or Pearson's r, neural activity regarding stimuli can be compared to each other, resulting in a representational (dis)similarity matrix (RSM or RDM)[161]. From there, non-parametric statistical tests can be conducted to compare neural activity across stimuli, groups or both [160]. In our case, we used a between-group analysis of anxious and non-anxious children, regardless of what stimuli they examined.

Traditional RSA has been widely adopted in cognitive neuroscience, but suffers from some confounding factors. Mainly, similarity metrics tend to be much higher when neural patterns are in close temporal proximity, which can conflate results [162]. Secondly, traditional RSA can result in unstable (bias) analysis when the signal-to-noise ratio is low for some sets of data [162], [163]. GBRSA is a Bayesian extension of RSA that can address the mentioned issues. While RSA uses deterministic approaches (e.g., general linear model or ordinary least squares) to estimate the similarity between the neural responses, GBRSA uses the maximum likelihood estimation (MLE) to learn hyper-parameters of a distribution for the neural responses of each subject — while a single covariance matrix is used across all subjects to maximize the joint probability of observing neural responses. we use a shared covariance matrix [163]. So, GBRSA improves on these issues by reducing the temporal and covariance bias — i.e., learning the covariance structure as a hyper-parameter. By reducing the unknown activity patterns across anxious and non-anxious children, a direct estimation can be made from the covariance matrix [162]. Once generated, we measured neural activity in the TP to determine whether either group differed from each other using a Mann-Whitney U-test, which does not assume our neural activity has a normal distribution. Analysis for GBRSA was done in Easy fMRI and Python [154], which included the library SciPy to conduct the

Mann-Whitney U-test [156]. Brain images seen in Figure 4.5 were done using Analysis of Functional NeuroImages (AFNI_21.1.01), Surface Mapping (SUMA) and Easy fMRI [151], [164], [165].

#### Fully connected network analysis

Lastly, we wanted to look at a fully connected network analysis between the highest selected region and all other regions. To do this, we first partitioned the raw neural activities between anxious and non-anxious children. Next, neural activities were further partitioned based on 959 regions of the Talairach atlas. We then averaged the neural activities within each Talairach region across all voxels — which resulted in a vector with the same size as our time points. After, we compared each of these vectors by calculating the absolute value of the correlation in a similarity matrix. We then applied a threshold to examine the most correlated regions (top 30%). Finally, we visualized both of the anxious and non-anxious networks and only showed the top connections with our highest selected region. This was done to examine whether our ROI showed different neural connections to different areas of the brain in anxious and non-anxious children, regardless of facial stimuli.

### 4.2.8 Analysis: Negative facial stimuli

#### Model classification of fearful versus angry faces

We also sought to distinguish fearful versus angry faces among the neural activity of all children with our ROI. For between-subject comparisons, tasks such as pattern classification or RSA yield lower accuracies because the representational spaces are highly dimensional, the functional topography may be different between subjects and anatomical brain structures vary between participants. Thus, a recent method known as functional alignment has been proposed to align patterns of neural responses across subjects into a common, lower-dimensional space [137]. One important assumption is that we assume all human brains have similar neural activity for experiencing the same categorical stimuli. Here, we used a probabilistic shared response model (SRM) to functionally align

neural activity for fearful and angry faces for all subjects only in our region of interest [166].

SRM uses the training data to learn the mappings for each subject's data shared feature space. Then these learned mappings are projected onto the held-out data for each subject into a shared feature space. One of the main distinctions in SRM is that the model directly estimates that the selected shared features are significantly less than the number of voxels it is selecting from. This is different from other methods, where the number of features usually equals the number of voxels [166]. The machine learning pipeline for our negative stimuli analysis can be seen in Figure 4.6. Analysis for SRM was conducted in Easy fMRI.

Once we obtained the functionally aligned dataset for the facial stimuli, we trained a linear SVM classifier on the ROI. Here, instead of using a non-linear model (such as AdaBoost in the primary analysis), we opted for a linear model instead. First, our intuition was that – since functional alignment maps the neural responses for our facial stimuli to a linear feature space, using a non-linear model would increase the chance of over-fitting. Thus, we followed Occam's razor and opted for a simple, linear model to prevent this issue. Secondly, using a linear model has reduced computational time compared to non-linear models. 5-fold CV was used, but with no internal CV approach this time. Also, no majority vote was used for final predictions. Instead, each time point was individually predicted, and metrics such as accuracy, precision, recall, and F1-score were averaged across each time point between all subjects in the testing folds.

**Multi-class classification**

Taking the final predictions from both our analysis models, we sought to make a four-class classification model that predicts the diagnostic stimuli (anxious versus non-anxious) and the type of stimuli (fearful versus angry faces) in a post-hoc analysis. Using the final predictions from the trained AdaBoost classifier (anxious versus non-anxious) and the linear SVM (fearful versus angry faces), we generated new prediction labels based on the original class and

stimuli labels and compared them to the observed labels. A one way ANOVA was conducted to examine differences in precision between the 4 classes. This can be seen in Figure 4.7C.

## 4.3 Results

### 4.3.1 Clinical and demographic statistical analysis

Table 4.1 shows demographic and clinical data from 22 anxious and 23 non-anxious children in our sample, including comorbidities and overlap between various anxiety disorders. When testing for group differences in age, a two-tailed t-test revealed a significant difference between the anxious and non-anxious group ($t_{(43)} = 2.03, p < 0.05$) and the SP cohort ($t_{(32)} = 2.36, p < 0.02$). When measuring functional impairment and emotional symptoms, the anxious groups differed significantly from the non-anxious group ($p < 0.005$). No statistical differences were found when comparing sex, ethnicity, handedness, IQ, or socioeconomic status between any of the groups. Note, we compare non-anxious individuals against each of the anxious subtypes for this statistical test only. Our machine learning task combines all anxious subtypes into the main anxious group, as shown in Table 4.1.

### 4.3.2 Anxious versus non-anxious classification analysis

#### Machine Learning analysis

Using the Talairach atlas (2mm), we use a super learner (SL) to segment the whole-brain data into 959 regions, then train a AdaBoost with the regions serving as a hyper-parameter. Here, the SL used nested cross validation (CV) (5-CV on the outer and inner loop) to partition and fine tune hyper-parameters. Figure 5.2 illustrates the machine learning pipeline for this section. The SL achieved the highest accuracy by using voxels from region #41 with 81% (STE +/- 1.46%) (MNI: x = 40, y = 11, z = -35) (Right temporal pole, right Cerebrum, Superior Temporal Gyrus, Brodmann area #38). When examining differences between the second and third ranked regions from our internal CV

process, the accuracy of region #41 was not statistically different from region #664, which had an accuracy of 77% (STE +/- 1.33%) (MNI: x = 10, y = -50, z = 20) (Right cerebrum, Limbic lobe. Posterior cingulate white matter) ($t_{(21)} = 0.16, p = 0.87$) or region #720, which had an accuracy of 76% (STE +/- 1.52%) (MNI: x = -52, y = -19, z = 7) (Left Cerebrum, Transverse temporal gyrus, Brodmann area 38) ($t_{(21)} = 0.18, p = 0.85$). This can be viewed in Figure 4.2, which includes the top 20 ranked regions from the SL's internal CV mean accuracy. Note that region #664 included white matter tracts that were proximal and inside region #41, and region #720 was the left hemisphere temporal pole.

Classification performance was measured using accuracy (percentage of correctly classified participants), precision, sensitivity (*i.e.* recall), and F1-score. Using our SL, we achieved an accuracy of 81% an overall precision of 80% as seen in Figure 4.7A, recall at 80% and an F1-score of 80%. Table S1 reveals the detailed results of our final SL model.

To ensure these results were not driven by several confounding factors, we observed individual accuracies across a number of variables including age, anxious subtypes, and the different fMRI scanner sites. First, we plotted the individual accuracy of every participant based on their class label in Figure 4.3. In Figure 4.4C, we can see the average accuracy for each class subtype. The control group had an average accuracy of 81%. Our model returned an average accuracy of 76% for children with both GAD and SP, 76% for SP and SAD, 83% for children GAD only, 70% for SP only, 94% for SAD and GAD, and 72% for SAD and SP. Of note, children with only SP (n = 3) had large variations between accuracies, so the 95% confidence interval (CI) is quite expansive. Figure 4.4A shows accuracies grouped by age. With region #41 alone, our AdaBoost model could perfectly predict children ages 9-10 while also maintaining accuracies about 75% for all other age groups. We also plotted accuracies of individuals based on scanner sites in Figure 4.4B to determine whether a given scanner was driving our model results. Although scanner #0 had half the participants compared to scanner #1, the accuracies between the two are almost identical (scanner #0 = 82%, 95%-CI (70.62 - 93.57), scanner

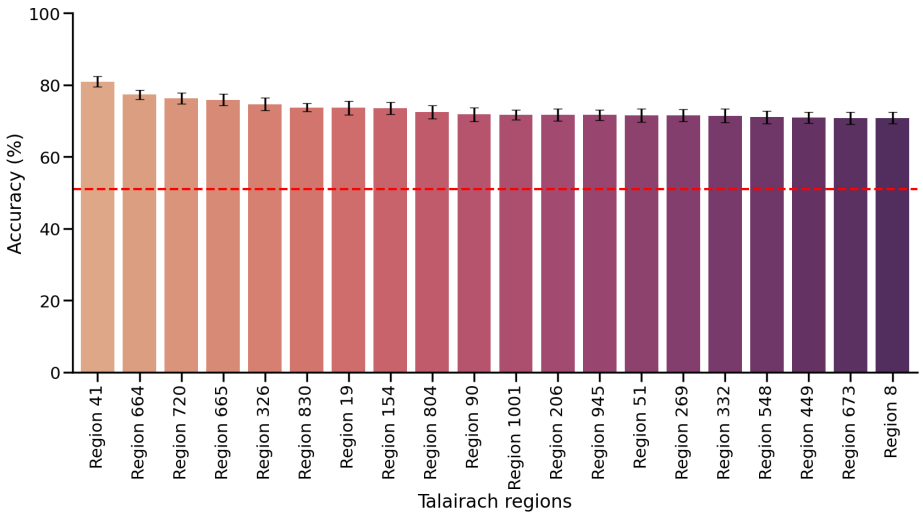



Figure 4.2: **Top 20 Talairach regions by accuracy (%).** The SL considered classifiers produced by base learners, each applied to a specific region. This shows the mean 5-fold internal CV accuracies, for the top 20 regions. The SL selected region # 41 to best distinguish anxious from non-anxious children. Other highly ranked regions include region # 664 (Right cerebrum, Limbic lobe. Posterior cingulate white matter) and region # 720 (Left Cerebrum, Transverse temporal gyrus, Brodmann area 38). All error bars represent the standard error of each participant across inner-CV folds from the SL. Red stripped line represents baseline accuracy for the majority group (51.1 %).

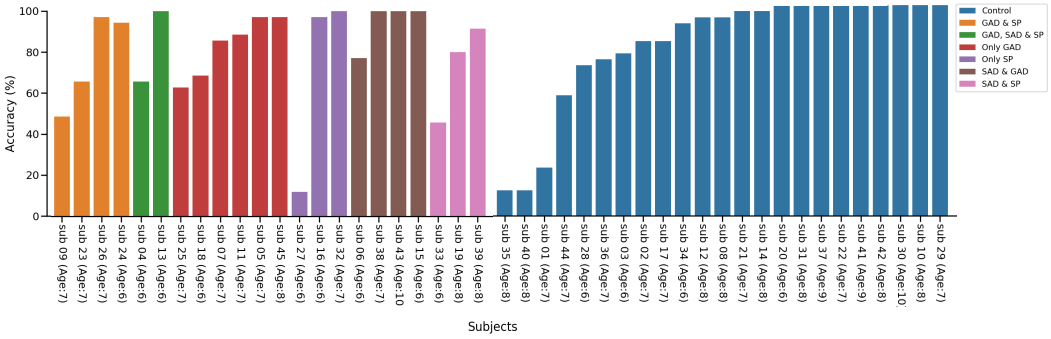



Figure 4.3: **Individual accuracies (%) of participants based on subclasses from final super learner model.** A plot of each participant's mean accuracy across 35 time points, grouped by subtypes from our final AdaBoost model, which learned from voxels in only region # 41.

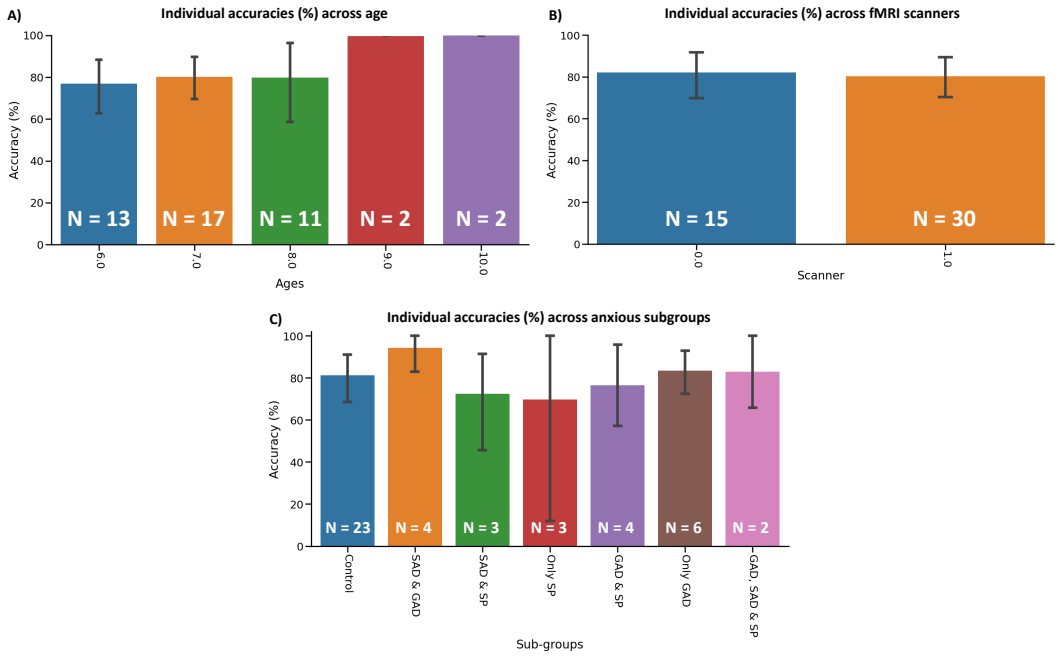



Figure 4.4: **A) Individual accuracies grouped across age (5-10) from super learner model**. Mean accuracy of individuals grouped by age. **B) Individual accuracies grouped across fMRI scanner sites**. Mean accuracy of individuals grouped by the two fMRI scanner sites. **C) Individual accuracies grouped by anxious subtypes**. Mean accuracy of individuals with one or more comorbid anxiety diagnosis. All error bars represent the 95% confidence intervals through bootstrapping of individual accuracies.

#1 = 80%, 95%-CI (70.24 - 90.45).

### Statistical analysis of Talairach region #41

Here, we conducted a high level, between-group ROI analysis for region #41 to examine the neural response differences between anxious and non-anxious children. Using the neural responses from our second level, grouped Bayesian representational similarity analysis (GBRSA), we compared activation in this region by using a mask to confine our analysis. Details on this method can be found in the online methods. This brain mask was used to extract this specific region only for our statistical test. Figure 4.5 (left) shows the region-based neural responses for both anxious and non-anxious children. We compared all 685 pairs of voxels in this region using a Mann-Whitney U test (two-tailed). The statistical test confirmed that the distribution of beta values from our

GBRSA analysis for the anxious group was significantly different from the non-anxious group in region #41 ($U = 215017.00, p < 0.005$).

Figure 4.5 (right) represents the fully connected network analysis for region #41. We examined neural activity of highly correlated brain regions with region #41 in both groups. We only drew connected regions with an absolute correlation threshold value of 0.6 or higher, and represented those using the red lines extending from region #41. Table 4.2 reveals 26 regions that have a correlation value of greater than 0.60 in the anxious group, but only 16 for the non-anxious group.

### 4.3.3 Negative stimuli classification

#### Model classification of fearful versus angry faces

As a posthoc analysis, we trained a linear Support Vector Machine (SVM) to predict whether a participant (either anxious or not) was viewing a fearful or angry face at a given time, using only voxels from region #41, which was predetermined from our SL. We applied a probabilistic shared response model (SRM) to functionally align the shared feature space across all subjects, as seen in Figure 4.6. Using 5-fold CV, with this new representational space, we achieved an accuracy of 97.1% (STE +/- 0.43%) with a precision of 97.5%, a recall of 97.1%, and an F1 score of 97.1% when classifying fearful and angry faces. Please refer to Table S2 for evaluation metrics. When we trained the same model without functional alignment, we only achieved an accuracy of 49.4% (STE +/- 0.81%), the precision of 45.1%, a mean recall of 49.4%, and an F1 score of 42.7% as seen in Table S2. Figure 4.7B reveals the precision of the SVM with functional alignment (SRM) and without functional alignment.

#### Four-class classification

This final classification model coupled the predictions from our primary analysis and negative stimuli classification models into a four-class performance task. For each time point, our ensemble model predicts which group the subject was from (anxious versus non-anxious), and what type of stimuli he/she was viewing at that time (anger versus fearful faces). Our model was able to

Table 4.1: **A summary table of demographic and clinical symptom scores of participants.** The anxious group contained individuals with one or more anxiety disorders from the mentioned subtypes. The 3 anxiety subtypes are not mutually exclusive. T-tests were conducted between the non-anxious group and all anxious subtypes, as well as the whole anxious group. Mean values and standard deviations are reported for all 4 groups. Significant difference from non-anxious children at $^{*}p < 0.05$, $^{**}p < 0.005$.

| | | **Non-anxious (N=23)** | ***Anxious (N=22)*** | ***Generalized Anxiety (N=15)*** | ***Separation Anxiety (N=10)*** | ***Social Phobia (N=11)*** |
|---|---|---|---|---|---|---|
| **Demographics** | Age at scan | 7.48 (1.04) | 6.86 (0.99)* | 6.86 (1.06) | 7.00 (1.33) | 6.63 (0.81)* |
| | Female | 13 | 16 | 12 | 7 | 8 |
| | Ethnicity | 12 | 10 | 8 | 6 | 3 |
| | Below poverty | 4 | 6 | 5 | 5 | 2 |
| | Handedness (right) | 16 | 18 | 14 | 7 | 8 |
| | IQ | 104.48 (14.02) | 103.86 (10.81) | 103.52 (11.51) | 103.20 (10.63) | 106.18 (9.54) |
| **Symptoms** | Impairment (0-10) | 0.74 (1.09) | 3.5 (2.35)** | 3.93 (2.66)** | 3.80 (2.62)** | 3.28 (1.68)** |
| | Emotional symptoms (0-14) | 2.17 (1.99) | 6.54 (2.91)** | 7.26 (3.13)** | 8.40 (2.91)** | 5.81 (2.40)** |

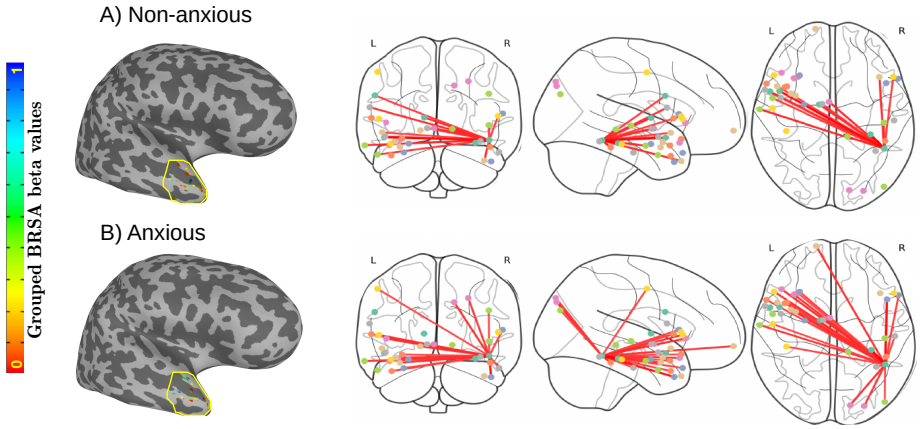



Figure 4.5: **Left) Grouped Bayesian representational similarity analysis of region #41.** A between-group ROI analysis was used to examine activation differences for anxious versus non-anxious children. In our statistical comparisons, 685 pairs of mean beta values in region #41 were compared between each group using a Mann-Whitney U-test (two-tailed). (U=215017.00, $p < 0.005$). **Right) Fully connected network analysis of Talairach region #41 with anxious versus non-anxious children.** A visual representation of regions connected with region #41. Dots represent the 38 regions with absolute correlation thresholds greater than 0.6, each connected with a red line to region #41. A) Fully connected network for non-anxious children. B) Fully connected network for anxious children.

Table 4.2: **Talairach regions most correlated with region #41.** These values represent the absolute Pearson correlation between different Talairach regions and region #41 (threshold to above 0.6). Twenty-six regions show a Pearson correlation with region #41 above 0.6 for anxious children, but only 16 for non-anxious children. Abbreviations: STG: Superior temporal gyrus, ITG: Inferior temporal gyrus, MTG: Medial temporal gyrus, WMT: White matter tract, IFG: Inferior frontal gyrus, VAN: Ventral anterior nucleus, SOG: Superior occipital gyrus, SFG: Superior frontal gyrus.

| **Brain Regions correlated with Region #41** | **Non-anxious** | **Anxious** |
|---|---|---|
| R. Temp. lobe, STG | 0.87 | 0.90 |
| L. STG, Brodmann 38 | 0.83 | <0.60 |
| R. ITG, Brodmann 21 | 0.65 | <0.60 |
| R. MTG, Brodmann 21 | 0.71 | <0.60 |
| L. STG | 0.74 | 0.75 |
| L. Fusiform Gyrus | <0.60 | 0.63 |
| L. Fusiform Gyrus, Brodmann 20 | <0.60 | 0.66 |
| L. Fusiform Gyrus, WMT | <0.60 | 0.66 |
| L. Parahippocampal Gyrus, Brodmann 36 | <0.60 | 0.66 |
| L. Fusiform Gyrus, Brodmann 36 | <0.60 | 0.68 |
| R. Amygdala | <0.60 | 0.67 |
| L. Front. lobe, IFG | <0.60 | 0.67 |
| R. Front. lobe, IFG | <0.60 | 0.83 |
| R. IFG Brodmann 47 | <0.60 | 0.72 |
| R. Font. lobe, IFG, WMT | <0.60 | 0.74 |
| R. Parahippocampal Gyrus, Brodmann 20 | <0.60 | 0.69 |
| R. Parahippocampal Gyrus, Brodmann 38 | 0.72 | <0.60 |
| R. Frontal lobe, STG | <0.60 | 0.72 |
| R. Temp. lobe, IFG | 0.73 | 0.71 |
| L. Sub-Gyral | 0.64 | <0.60 |
| R. Temp. Sub-Gyral Brodmann 13 | 0.66 | <0.60 |
| R. Temp. Sub-Gyral | 0.73 | <0.60 |
| R. Front. Lobe, Sub-Gyral Brodmann 47 | <0.60 | 0.63 |
| R. Brainstem Extra-Nuclear WMT | <0.60 | 0.64 |
| R. Cerebrum Extra-Nuclear WMT | <0.60 | 0.66 |
| L. Temp. Lobe, Insula Brodmann 13 | 0.64 | <0.60 |
| L. Front. Lobe, IFG Brodmann 45 | <0.60 | 0.73 |
| R. Front. Lobe, IFG Brodmann 45 | <0.60 | 0.74 |
| R. Front. lobe, SFG | 0.71 | <0.60 |
| R. Thalamus, VAN | 0.64 | <0.60 |
| R. Front. lobe, Precentral Gyrus | <0.60 | 0.70 |
| R. Front. Lobe, IFG Brodmann 44 | <0.60 | 0.71 |
| Cuneus | 0.70 | <0.60 |
| L. Occip. Lobe, Ceneus Brodmann 19 | 0.65 | <0.60 |
| L. Temp. Lobe, SOG | 0.66 | <0.60 |
| R. Front. Lobe, IFG, Brodmann 9 | <0.60 | 0.63 |
| L. Occip. Lobe, SOG Brodmann 39 | <0.60 | 0.64 |

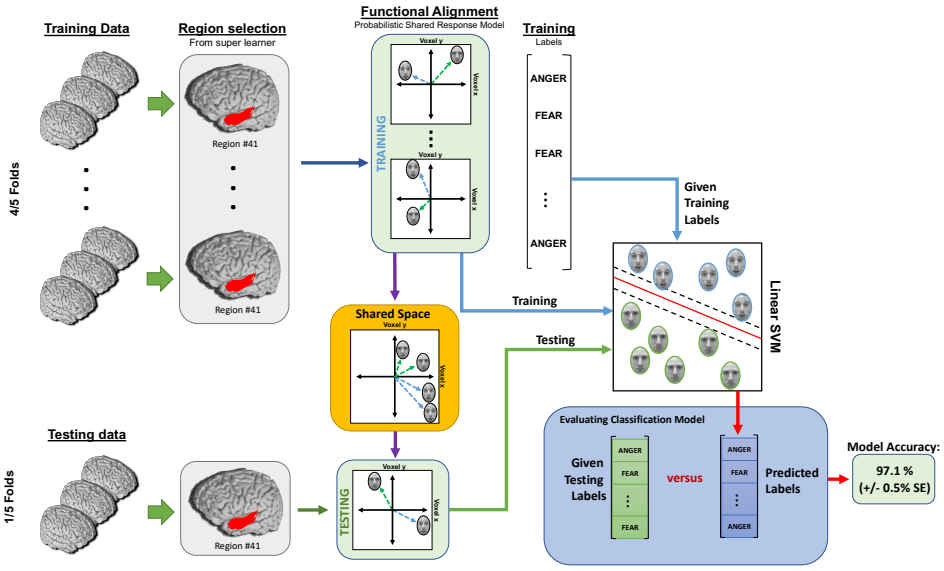



Figure 4.6: **Posthoc analysis: Preprocessing pipeline for negative stimuli analysis**. Voxels from the selected region (from the primary analysis) were used to predict the stimulus label for each time point (fear versus anger). We used a probabilistic shared response model (SRM) to transform all functional images into a shared common space. Thereafter, a linear SVM was trained on the functionally aligned data, which were used to predict a facial stimulus for each time point, for each subject. Model metrics include mean accuracy and standard error from 5-fold CV.

achieve a balanced accuracy of 73% (STE +/- 0.06%) which is an improvement from baseline (26%). Mean precision, recall, and F1 scores were also 73%. Non-anxious children viewing angry faces revealed the highest recall at 76%, and non-anxious children viewing fearful faces revealed the highest precision at 75% as seen in Figure 4.7C and Table S3. No significant differences were found between the 4 classes with respect to their precision, recall, or F1 scores.

## 4.4 Discussion

This study illustrated that a data-driven, machine learning approach can be used to distinguish anxious children from non-anxious children and identify which regions may be important for this performance task. Talairach region #41 (aka, Brodmann's area 38, right temporal pole, planum polare, or area TG)

can be used to distinguish anxious children from non-anxious children based on their brain scans, as they view negative facial stimuli. We demonstrate that task-based fMRI activity related to this anatomical area is sufficient to achieve a relatively high accuracy. In our primary analysis, we trained an SL to parcellate regions and train a non-linear model on those regions. Region #41 was selected as part of our final model, which had an accuracy of 81%. This was compared to a model which included whole-brain activity (54% accuracy), and a searchlight analysis (59.2% accuracy). When examining confounding variables such as age, anxious subtypes and scanner sites, they did not seem to drive the accuracy of our model. Though, we could not run statistical tests because the subtypes and age groups samples were too small in some cases. It is difficult to discern whether our model is partial to certain subtypes of anxiety, but our results suggest the SL performed well regardless. Additionally, we examined the functional connectivity between region #41 and other areas and found that anxious children showed similar correlational patterns between several regions that make up the affective network. In addition, anxious children also exhibited more and stronger correlational patterns to other brain regions compared to non-anxious children. This network is a distributed neuronal network related to mood regulation and affective processing [139], [167]. However, very little research has been conducted on this particular region in relation to pediatric anxiety.

In our posthoc analysis, we examined how neural signatures differed between fearful and angry faces in both anxious and non-anxious children for region #41. We were able to achieve an accuracy of 97% with a linear SVM, but only after applying functional alignment (probabilistic SRM) to the brain scans of all children. This suggests that fearful and angry faces are highly dissociable when projected onto a common shared space. Functional alignment can provide enhanced predictive power because it automatically reduces the feature space while aligning the vectors between subjects to a shared common representational space [137]. We then trained a linear model to make individual predictions from both of our previous models in a four-class classification task that predicted the disease state and the type of facial stimuli simultaneously.

Here, we achieved an accuracy of 73%, suggesting we can identify both neural signatures of anxious children and how they process fear and angry faces.

Due to the diverse structure and connectivity to a number of regions, the putative role of the TP has been inconsistent and subject to significant debate [122], [168], [169]. The TP has been proposed as a social-emotional cognition hub that receives various sensory inputs from limbic structures to organize social processes [168]. Emotional facial processing is a particular social process, and young children suffering from anxiety seem to show functional dysregulation in related key limbic structures such as the amygdala and the PFC [127], [131]. Our results show that the TP also plays a crucial role in facial processing in children. Using only the neural correlates in the TP, we were able to make individualized predictions about which children suffered from anxiety using a non-linear machine learning model. This suggests that altered functionality exists in this region during a facial processing task involving negative or threat provoking stimuli. This is a novel finding in relation to pediatric cases of anxiety.

### 4.4.1 Neuroanatomy of the Temporal Pole (TP)

The TP lies between the O-PFC and the amygdala [122], sitting near the anterior end of the temporal lobe, rostral to the perirhinal cortex. It has significant neural connections with the amygdala and PFC via the uncinate fasciculus, making it a paralimbic region [122], [170]. Although it is known for processing language, functionality surrounding the TP has also been linked to facial, emotional, and social processing, but it still remains largely understudied [171]. Below, we present findings in neuroanatomical studies in macaque monkeys, showing patterns of connectivity similar to those exhibited in human brain [172]–[174]. Additionally, in the human brain, the TP is anatomically close to (and highly connected to) other areas related to facial and socioemotional processing [122].

### 4.4.2 Evidence of socioemotional processing in the temporal pole

Using data from macaque monkeys, Kondo, Saleem & Price (2005) hypothesized that the TP modulates emotional functions related to salient perceptual stimuli based on anatomical connectivity [175], [176]. The ventral region receives input from visual processing centers and is considered to be an endpoint in visual processing in macaques [175]. Neurons in this region respond to complex stimuli and change in activity related to visual memory tasks [172]. In humans, neuroimaging tasks have shown that neurons in the TP also responds to complex visual stimuli such as faces. Additionally, studies of visually evoked negative emotions, such as fear and anger, have observed changes in activity of the right-ventral region of TP [177], [178]. Right-lateralized regions of TP have been implicated in high-level sensory representations with emotional and social experiences, while left-lateralized regions of TP have been associated with linking semantic memory to high-level representations such as faces [122]. Specifically, damaged left TP studies revealed deficits in proper naming abilities and face-name associative learning tasks [179], [180]. Additionally, epileptic damage to the right TP has resulted in higher prevalence of anxiety and depression disorders compared to the left TP [179]. Here, it is evident that the TP plays a role in processing emotionally balanced facial expressions. In childhood anxiety, this cognitive process is compromised. Research remains focused largely on amygdalocentric systems, but our results suggest that the activity of right TP alone enables distinctions between anxious and non-anxious children.

As outlined above, research into neural aspects of socioemotional processing has focused mainly on areas such as the amygdala and the prefrontal cortex, not the TP. It has not received the same attention as the amygdala and PFC in emotion studies. [122]. In addition, the awkward anatomical placement of TP near the air-tissue boundaries of the sinuses is associated with weaker BOLD signals, making it difficult to draw consistent and statistical significance from fMRI measures [181]. Nevertheless, in our study, the right TP revealed

neural differences in socioemotional processing of facial stimuli in anxious and non-anxious children.

### 4.4.3 Implications in Childhood anxiety

Since 2000, a number of neuroimaging studies have found deficiencies in facial-emotion recognition among individuals suffering from anxiety. In most cases, research has focused on two brain regions — amygdala and PFC — when examining distinct functional differences in anxious individuals while viewing emotional faces [179]–[182]. Cognitive schema theories suggest that negative or threatening faces receive preferential and early processing advantages through these parts of the brain. It is known that rapid, direct processing of rudimentary sensory stimuli from the thalamus can reach the amygdala in short succession [183], [184], and the amygdala also receives sensory input from indirect pathways such as the PFC, where it assigns significance to the sensory stimuli based on context and prior experiences. This pathway reaches the amygdala in a slower fashion, but conveys higher-order representations and is relevant for memory consolidation [185]. Top-down modulation of this pathway may exert inhibitory influences on the amygdala. In adolescent anxiety, modulatory pathways may become dysregulated, allowing the amygdala to become hyperactive. Numerous studies have cited the PFC-amygdala network in emotional facial processing tasks [181], [183], [186], [187]. While this area is relatively well-known to be implicated in childhood anxiety, it is not the only area where functional differences may be observed.

In trying to conceptualize the functional relationship between anxiety and facial emotional processing, the TP should be considered. American neuroscientist Joeseph LeDoux argued the notion that the amygdala is the fear center of the brain. Instead, he posited that conscious fear is a cognitively assembled experience, derived from many other brain regions, which is not to be confused with the amygdalocentric, non-conscious process of detecting and responding to threats [188]. Additionally, several theories on emotional processing have moved away from the notion that the amygdala is part of a dedicated and prioritized network for assigning emotional valance to ecologically salient stim-

uli [189]. Revised hypotheses posit that the amygdala is a modulatory center with wide-ranging networks to other brain areas. Thus, it may be responsible for processing information related to salience, significance, ambiguity and uncertainty, assigning biological value to external stimuli [189]–[191]. Others have cited the amygdala as part of a "whole-brain phenomena" for constructed emotions [192]. One author postulated that the dynamics of the amygdala is used to make predictions of the external world, rather than react to it. This is done in the service of allostasis (an organisms' attempt to efficiently ensure resources for physiological systems in order to survive and reproduce) [192], [193]. Emotions serve as constructions or predictions of the external world. They are part of an integrated system that mobilizes the brain and the body to ensure allostasis is maintained. Thus, the amygdalocentric view of emotional regulation should be revised to include a wider array of brain regions, such as the TP.

Few studies have focused on paralimbic regions such as the TP and its connectivity to the amygdala, perhaps due to its later, high-level processing onset of complex stimuli. As mentioned earlier, the TP is (1) heavily connected with the amygdala, (2) responsible for processing complex visual and auditory stimuli, and (3) has been shown to integrate social and emotional significance to said stimuli. Together, these areas are a part of a larger-scale, resting-state brain network known as the affective network [167]. This network has been implicated in various anxiety disorders where individuals are characterized by hyper-arousal, heightened worry states, increased sensory processing and poor emotional regulation [140]. Our connectivity analysis provides strong evidence that the affective network is at play in adolescent children. Not only was the TP a strong predictor of childhood anxiety, but it also showed strong correlational patterns with other affective network regions such as the amygdala, STG, the orbital frontal cortex (OFC), and the border of the insula (Brodmann area #45) as seen in Figure 4.5 (right) and Table 4.2. In addition, anxious individuals exhibited strong correlations for neural activity between the TP and visual cortical areas such as the occipital cortex and the fusiform gyrus. These regions have been linked to the perception of emotion in facial

stimuli [194]. Although the affective network domain is based on resting state brain paradigms, similar regions in our task based study were identified by our model. Regions within the affective network seem to show the greatest discriminative power between anxious and non-anxious children. This only reinforces the notion that the TP must be examined in more detail, as it is part of a larger affective network that regulates emotional and perceptual stimuli.

Various brain studies have focused on functional and structural differences in the bilateral TP between different populations during emotional stimulus tasks. While none have emulated the paradigm we have presented, certain parallels exist to confirm our findings. From a social and emotional standpoint, some fMRI and PET studies have shown activation of the TP in such tasks. Bilateral activation of the TP has been seen in negatively valenced films inducing sadness [195], [196], and right TP activation has also been noted in viewing sad and angry faces [197]. Recalling past anxious and angry experiences have shown bilateral activation as well [198]. This suggests that emotionally valenced stimuli do operate within the TP. In our study, we applied functional alignment techniques to better distinguish between various emotional stimuli such as fear and anger in anxious and non-anxious children. Our goal here was to further examine whether types of emotional stimuli differ in their neural signatures for anxious and non-anxious children in the TP. Using only this region, we were able to distinguish diagnostic labels using the emotional stimuli of fearful and angry faces. This implies that anxious children process fearful and angry faces differently from each other, and they also process these emotions differently from non-anxious children.

While few studies have been conducted on children with anxiety, there are a few sources that cite the TP in clinical anxiety. A study focusing on group-level differences between SP and GAD in young adults with SP showed increased BOLD activity in the TP and the amygdala when responding to fearful faces compared to young adults with GAD and the control group [199]. The GAD group showed increased activity for angry faces in Brodmann area #10 (which includes part of the PFC) and the middle frontal gyrus, but showed a decrease in activity for the amygdala compared to SP and control

groups. The authors concluded that the amygdala was not a sufficient area alone to distinguish group and stimulus level differences in young adults with SP and GAD, recommending that other areas be examined as well [199]. A meta-analysis on SAD in adults revealed abberant activity in affective and default mode network regions such as the right TP, insula, PFC and the precuneus. The authors provided evidence that cognitive processes such as self-referential processing and Theory of mind are linked to these brain regions in SAD adults [200]. Lastly, a resting-state MRI study that analyzed the functional connectivity between the amygdala and the TP in GAD patients revealed some interesting caveats [201]. Compared to the control group, GAD patients revealed an increase in functional connectivity between these regions. They contend that this altered connection may contribute to the etiology of GAD in older patients. Our study confirms the same findings, except for children. Based on our connectome analysis, anxious children had a higher correlation between the right TP and the amygdala as well as the PFC, while non-anxious children did not. This may reveal that innervation's between the TP and other limbic regions may manifest in young children, and remain into adulthood.

One looming question that remains is the progression of anxiety and its relation to TP as an individual enters adulthood. Specifically, how does the TP affect social and cognitive abilities for a child entering adulthood with clinical anxiety? One premise to consider is Theory of mind and how dysregulated socioemotional processing can result in a failure to respond adequately to social interactions [122]. The inability to properly assess emotional faces and information of others could result in reduced positive and rewarding emotions after a social interaction, leading to maladaptive behavior [202]. A review on SAD in adults showed strong correlations between abberant self-referential processing, Theory of mind and subsequent dysfunction in sub-cortical brain regions such as the TP [200]. Additionally, adults with damage to the right TP have exhibited introversion and coldness, perhaps due to the failure to derive pleasure from social interactions [203]. Reinforced behavior to partake in socialization may be reduced and persist into adulthood [204]–[206]. This

trend has been seen in disorders such as Attention-deficit hyperactivity disorder (ADHD) and autism (ASD) [207], [208]. In non-clinical groups, children who benefit from accurate cognitive reappraisal and theory of mind go on to show linear, or even quadratic increases in activity in areas such as the right TP as they move to adulthood [202]. While it is difficult to ascertain if the TP directly causes socioemotional dysregulation into adulthood for anxiety, the fallout from poor theory of mind may persist into adulthood. Anxiety disorders are among the most persistent mental health disorders propagating from adolescence to adulthood, with a core criteria of symptoms centered around social and emotional processing [209]. It may be that dysregulation of the right TP in childhood anxiety may proliferate into adulthood.

Currently, there is no validation or diagnostic procedure that involves any component other than clinical signs and symptoms via psychiatric assessment. Although neuroendocrine, cognitive, genetic, and neuroanatomical correlates exist, there is no available biological test for diagnosis. Here, we used a facial processing fMRI task to not only classify anxious from non-anxious children, but to also distinguish between the affective stimuli presented. Instead of using a multi-variate strategy to examine whole-brain neural patterns, we focused on one particular region, which has been implicated but understudied in anxiety and socio-emotional processing. The TP served as an anatomical region that could predict which children suffered from anxiety based on the neural correlates of fearful and angry faces.

### 4.4.4 Limitations and future work

One future consideration is to conduct a meta-analysis to other studies of anxious children or adults. As mentioned, there are few papers focusing on the temporal pole as a region that could be implicated in anxiety. If data is available, validating our model on adult cases could yield interesting findings. Another consideration is to test our model on a separate cohort dataset with more subjects. Ideally, evaluating the performance of our proposed approach could benefit from a more homogeneous target group. Since our target group contained children with 3 different types of anxiety disorders

(co-morbid disorders as well), the variance between the neural signatures of these children may have differed significantly. There is evidence that SP, GAD, and SAD show different neurological and behavioral patterns between each other [145], [199], [210]. Thus, confounding effects may exist within the anxious group that could affect the overall performance accuracy. However, comparing SP, GAD, and SAD is out of the scope of this paper. Another limitation may be that functional alignment could result in high accuracies in other areas than the TP for our secondary analysis. The TP was critical for the primary analysis classification, but was comparable to other brain regions in the secondary analysis after functional alignment was applied. Another future consideration could involve transfer learning, a machine learning technique that involves training with one type of labeled data (*e.g.* only train a model using GAD participants), then applying that model on other classes (*e.g.* SP or SAD children) to examine whether the model can correctly distinguish cases based on the prior knowledge of only GAD children. Another potential limitation was the absence of other facial stimuli in the task. Carpenter et al., (2015) only released functional scans with fearful and angry faces for the purposes of their study [145]. Thus, we had to focus on negative stimuli only, and although we successfully distinguished fear from angry faces in the brain, other facial stimuli may offer further insights into how emotion is processed in the brain of anxious and non-anxious children. We may extend our study to the other related, publicly available datasets that have other types of (visual/facial) stimuli.

## 4.5 Conclusion

In summary, the goal of this study was to use a data-driven approach to classify anxious versus non-anxious children using emotional facial stimuli. Here, we used a super learner (AdaBoost with logistic regression as a base estimator) to select the Talairach regions that could best distinguish anxious from non-anxious children. Our model achieved an accuracy above 81% for this task. Subsequently, we examined how different negative emotional faces would be processed in both groups. We found that fear and angry faces could clearly

be distinguished in the TP, but only after functional alignment was applied to the brain scans of all subjects. This study illustrates the power of task-based fMRI designs to predict disease states and stimulus conditions. It also indicates that the TP is a region that should be further examined in pediatric anxiety. Cognitive processes such as emotional facial processing may be compromised in anxious children. We have demonstrated that machine learning analysis of face-processing, task-related fMRI data may be used to distinguish anxious from non-anxious children. This may enable further understanding of neural underpinnings of pediatric anxiety and help to extend and validate diagnostic labels used by psychiatrists and other clinicians.

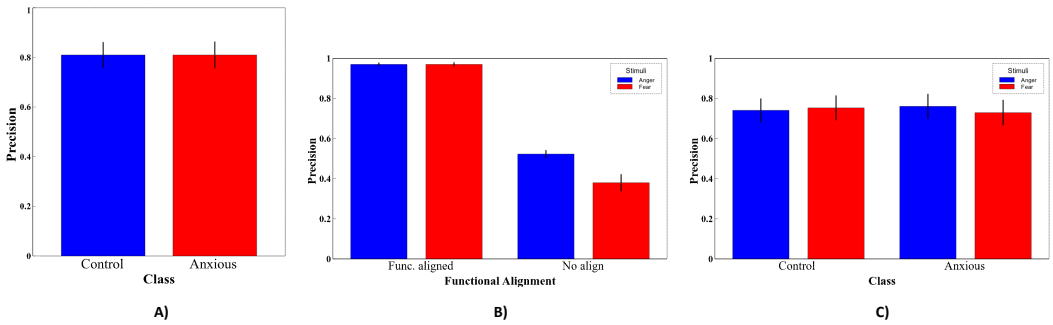



Figure 4.7: **Precision tables for both anxious versus non-anxious and negative stimuli classification analysis.** **A**) Precision table for anxious versus non-anxious AdaBoost classification model. **B**) Precision table for fearful versus angry faces, linear SVM with and without functional alignment. **C**) Precision table for four-class classification model of negative stimuli and disease state. All error bars represent the standard error of the precision across outer-CV folds.

# Chapter 5

# Detecting presence of PTSD using sentiment analysis from text data

## 5.1 Introduction

Post-traumatic stress disorder (PTSD) is a debilitating condition initiated by exposure to traumatic events, whether witnessing the event in-person, indirectly learning that a traumatic event occurred to a loved one, or through repeated exposure to aversive details of said events [211]. There is strong evidence that the current severe acute respiratory syndrome coronavirus two (SARS-CoV-2) pandemic can be considered a global traumatic event. [212]. Two outcomes have emerged from the SARS-CoV-2 pandemic: (1) There has been a surge in stress-related mental illnesses such as PTSD, specifically in occupational settings [213]–[217], and (2) many in-person medical appointments have been moved to a digital format [218]. Although only a small percentage of individuals develop PTSD following a traumatic event [219], the current pandemic has exposed distressful situations among many. Prior to the global pandemic, a general population survey across 24 countries estimated that 70% of individuals would experience at least one potentially traumatic event (PTE) in their lifetime [211]. That figure is now estimated to be higher due to the SARS-CoV-2 outbreak [220]. Approximately three in every ten survivors of the SARS-CoV-2 virus, two in every ten healthcare workers, and one in every ten individuals of the general population have reported an official diagnosis of

PTSD or PTSD-like symptoms [220]. One study estimated that roughly 25% of the general population in the United States has suffered from PTSD during the pandemic [221]. With PTSD prevalence rising, it is imperative that we improve on screening and diagnosis, especially with the current emergence of telepsychiatry.

Machine learning (ML) provides a computational tool to better understand the emotional and behavioral nature of PTSD, by learning general rules and patterns from large amounts of patient data [222]. With the increasing rate of online assessments, automated identification of disorders such as PTSD can be a useful tool of e-health. Much work has focused on learning a diagnostic classification model that can answer: "Does patient X have PTSD?". Here, a learning algorithm uses a set of labeled instances, to produce a model that uses information about a novel patient (perhaps blood factors, functional neuroimaging, speech and text data) to predict a 'label' value (perhaps 'Yes' or 'No') [223]. Once trained, a model can make predictions about a novel instance, which we hope are accurate.

Currently, diagnosis of PTSD is done through a clinical interview, which can be inaccurate due to subjective assessments and expertise bias. For example, PTSD is often under-diagnosed and conflated with more prominent disorders such as depression [224], which can affect prognostic outlooks. Second, the etiology of PTSD is multi-causal and complex. Due to the multi-faceted nature of trauma and its kaleidoscopic impact on individuals, clinicians are left to sift through heterogeneous phenotypic expressions [225]. This is problematic as: (1) Heterogeneous phenotypic expressions make it arduous for clinicians to assess and treat symptoms of PTSD and (2) differentiating between PTSD and other conditions may be difficult. Next, individuals may feign a PTSD diagnosis for several reasons including legal, personal, social or financial issues [226]. Finally, clinicians also face adversity when calling into question the validity of self-reported trauma or related symptoms, as they may worry about stigmatizing patients or losing rapport with potential victims of trauma [227]. Despite the complications, accurate diagnosis can provide patients with adequate treatment earlier, and it can also allow for healthcare systems to properly allocate their

resources to those who need it most.

Natural language processing (NLP) is a branch of artificial intelligence that handles text data to decipher and understand human language and context [228]. As discussed in Section 2, machine learned models have been applied to text data to accurately identify and treat individuals with, or at risk of, developing PTSD. Natural language has several advantages over other types of modalities, such as brain imaging, metabolomics or genomics: It can be collected at low cost, requiring no more than an audio call, it directly expresses emotions and thoughts through content, it is non-invasive, and it is difficult to conceal or feign symptoms [229]. Linguistic content may reveal significant information about an individuals' internal state. Speech is a complex form of communication that interweaves expressive thought, emotion and intention. It is a window into the mind, and can serve to detect markers of psychiatric illnesses [229]–[232]. Our study uses sentiment analysis (a sub-branch of NLP) to gauge the emotional valence of textual data from individuals suffering from PTSD. Our task is to predict whether an individual may be suffering from PTSD using the emotional valence of their text data in a conversational interview.

In this paper, we use sentiment analysis techniques to detect the presence of PTSD, using text data from a popular dataset, the Audio/Visual Emotion Challenge and Workshop (AVEC 2019) [233], which is a subset of the larger Distress Analysis Interview Corpus of Human and Computer interviews (DAIC_WoZ) [234]. The DAIC_WoZ is a multi-modal dataset containing recordings (audio and visual) and transcripts from semi-structured clinical interviews with individuals suffering from PTSD and/or major depressive disorder (MDD), as well as age-and sex-matched controls [234]. The protocol was designed to identify people with such disorders. Interviews are conducted by a virtual agent (Ellie) presented on a television screen. Ellie is controlled by a human operator to ask a series of questions to the participants. Two sets of psychiatric questionnaires were used to assess levels of MDD and PTSD [235]: the PTSD Checklist-Civilian version (PCL-C) and the Patient Health Questionnaire-Depression 9 (PHQ-9). Our performance task is to predict which individuals are suffering from PTSD using the emotional valence from the transcripts pro-

vided. The result of the PCL-C questionnaire serves as our outcome variable in this study. This dataset has been examined by previous researchers, who used machine learning tools to better predict which individuals had MDD. To our knowledge, we have not found any studies which only use text data to predict PTSD in these individuals. Rather, previous studies such as DeVault *et al.* (2013) and Stratou *et al.* (2013) incorporated audio, motion tracking and text data to predict PTSD. We want to illustrate that emotional language alone can be used to predict PTSD in these individuals, something which has not been done on this popular dataset. However, like those other studies, we plan on incorporating audio and motion tracking data afterwards in a separate study. Thus, the goal of this study is to illustrate that our sentiment analysis can provide accurate predictions, while only using text data on the AVEC-19 dataset, something which has not been accomplished before. We believe this is an important endeavor with the ongoing pandemic and the mental health epidemic happening right now. We also propose that our simple set of features can be used in conjunction with other types of data to improve upon diagnostic accuracy. This may be part of future studies.

Section 2 covers related works using NLP approaches to predict the presence of PTSD in individuals. Section 3 then describes sample demographics and walks through our methods, which include the sentiment analysis pipeline, feature engineering, and the learning procedure. Section 4 discusses the results of our analysis. Lastly, we discuss the space for linguistic analysis in PTSD, and how language can serve as a primary indicator for measuring symptoms.

## 5.2 Related Work

In recent years, there has been growing interest in building automated systems that could screen for PTSD in individuals. Some approaches involve learned models that use text mining or NLP approaches. A text-based screening tool involves multiple components, including data acquisition (an audio recording of an individual's responses to a set of designed questions), feature extraction (quantifying features generated from the dialogue, such as sentiment analysis,

bag-of-words or word embeddings), and classification, which applies a trained ML model to those verbal features in order to predict whether a new patient is suffering from PTSD or not [238]. He *et al.* (2012) used lexical features in self-narratives from 300 online testimonies of individuals with PTSD and a control group. Their ML pipeline represented their verbal features using 'bag-of-words', which counts the number of occurrence of keywords in a given document. After, they trained a chi-square model [240] (a method for document classification based on using the chi-square test to identify characteristic vocabulary of document classes) and achieved an accuracy of 82%. He *et al.* (2017) conducted another study examining self-referential narratives about traumatic experiences in a clinical screening process. In that study, they used 'N-gram' features (which count the number of co-occurring words within a given window of words). Their Product Score Model with a uni-gram feature space attained an accuracy of 82% over a corpus with 300 individuals (150 with PTSD) who filled out an online survey related to their mental health; this is the highest of all algorithms they tested.

Some feature engineering methods examine the emotion or sentiment of textual data, to produce features that can be used as indicators of several symptoms [242]. Language through media such as social media can convey feelings of negativity towards one's self. Large datasets can be derived from social media platforms, which can be useful in training models that can generalize to a wide range of individuals suffering from disorders such as MDD or PTSD. De Choudhury *et al.* (2013) learned a probabilistic model that could detect depression based on Twitter data. The authors generated features based on the emotional sentiment of those tweets, then used dimensionality reduction methods (Principal Component Analysis [PCA]). Using a Support Vector Machine (SVM), they achieved a 74% accuracy in detecting depression. Another study used a pre-trained language analysis tool called Language Inquiry Word Count (LIWC) to extract the emotional polarity of sentences into an overall score. Sentences with the words 'mad', 'sad', 'fail', 'cry' returned a more negative compound score, while words such as 'happy', 'joy' and 'smile' returned a positive compound score [244]. Though this was not an ML study, they did

Figure 5.1: **Interview process with Ellie.** Participants were placed in a room in front of a large computer screen, showing the animated character Ellie in the Wizard-of-Oz interview.

find significant differences in the linguistic style of individuals suffering from emotional distress compared to those who were not. Another language analysis tool, VADER (Valence Aware Dictionary for sEntiment Reasoning), is a rule-based model that uses both qualitative and quantitative methods to determine the sentiment intensity when humans are verbalizing [245]. VADER improves on LIWC by containing a larger word corpus and by being less computationally expensive and more easily implemented. Leiva and Freire (2017) used VADER to predict whether someone is at risk of developing depression using sequential social media messages. They reported that VADER was the best sentiment analysis method for predicting whether a user is at risk of depression or not based on the Early Risk Detection Score (ERDS) [242]. Sentiment analysis is a useful and simple method to implement on a corpus of text data. Yet, there has been little research done on predicting PTSD using sentiment analysis on an individual basis.

We apply sentiment analysis algorithms such as VADER to the AVEC-19 dataset to determine whether emotional intensity of interview transcripts can serve as predictors of PTSD or not.

## 5.3 Methods

The Distress Analysis Interview Corpus (DAIC) is a large, multi-modal database of semi-structured interviews [234], [246]. The original project was started at the University of Southern California (USC), and was approved by the USC Review board (UP-11-00342). The current study includes a secondary analysis of the DAIC dataset, which was designed and collected by Gratch *et al.* (2014) at USC. Our study, which was approved by the University of Alberta's Health Research Ethics Board (Pro 00072946), is a secondary analysis, that did not involve collecting the data nor designing the study. We provide an extensive account of their documented methods below. Individuals with PTSD or MDD participated in a virtual clinical interview with an artificial avatar named Ellie [234]. This study was done to compare the development of computer-assisted rates of diagnosis with human performance [234]. Ellie was controlled by a human, who administered a series of questions to the individual in a semi-structured manner, while responses were recorded and transcribed to text. Figure 5.1 reveals the set-up, showing the automated interview with a participant and Ellie. In addition to text, their audio sample was collected, as well as their motion and eye tracking. These interviews were part of a larger project called *SimSensei*, which is developing virtual agents that interview individuals with mental health problems. They are using verbal and nonverbal indicators to screen for cognitive or behavioral abnormalities related to several illnesses [235], [247]. Our study only examined the transcribed text data, as our primary focus was to examine whether text data alone could detect the presence of PTSD.

### 5.3.1 Participants

The DAIC_WoZ dataset includes participants from two populations: U.S. armed forces military veterans (recruited from the U.S. Veterans Facility in Long Beach, New York) and civilians [234], [246], recruited from Craigslist [234], [235]. In the DAIC_WoZ subset, participants were flown into the USC Institute for Creative Technologies to participate on-site, in front of a TV. Target

participants between the ages of 18-65, who previously had a diagnosis of MDD or PTSD. All participants were fluent English speakers, and all interviews were conducted in English [234]. The sample included 275 individuals (105 females, 170 males), 188 controls and 87 who met the criteria for PTSD. Some of those within the PTSD group also met the criteria for MDD. The PTSD Checklist-Civilian version (PCL-C) and the Patient Health Questionnaire-Depression 9 (PHQ-9) were used as our outcome metrics [235]. We conducted a chi-square test to determine whether if the sex ratio was significantly different between the two groups. We also conducted a two-sample t-test for testing the mean PCL-C and PHQ-9 scores between the two groups. Note that the PCL-C is not used for official diagnosis of PTSD, but it is strongly correlated with the Clinician Administered PTSD Scale (CAPS-5), which is the gold standard measurement for diagnostic efficacy [248]. Table 1 illustrates the symptoms scores for both of these questionnaires across the control and target groups.

### 5.3.2 Procedure

Prior to the recorded interview, participants were given an explanation of the study, and then voluntarily signed a consent form [234]. Then, a series of questionnaires were conducted online, which included a demographics section, the PCL-C and the PHQ-9.

After completing the questionnaires, participants sat in front of a virtual character (Ellie), who was projected on a 50-inch T.V. monitor [234] as seen in Figure 5.1. Participants were recorded on a Logitech 720p webcam, and used a Sennheiser HSP 4-EW-3 microphone [234], [235], audio recording at 16 kHz. Acoustic data was recorded and stored by SimSensei [247]. For text data collection, SimSensei used the PocketSphinx recognizer to recognize spoken words for Ellie and the participants, and saved them in a document [235], [247], [249]. The individuals controlling Ellie used the Flores Dialogue Manager to decide on the proper responses and questions to ask the participants [250]. Ellie first explained the purpose of the study, then asked a series of 'ice-breaker' questions to build rapport with the participants [234], [235]. It then asked a series of emotionally valenced questions, such as: 'What are some things

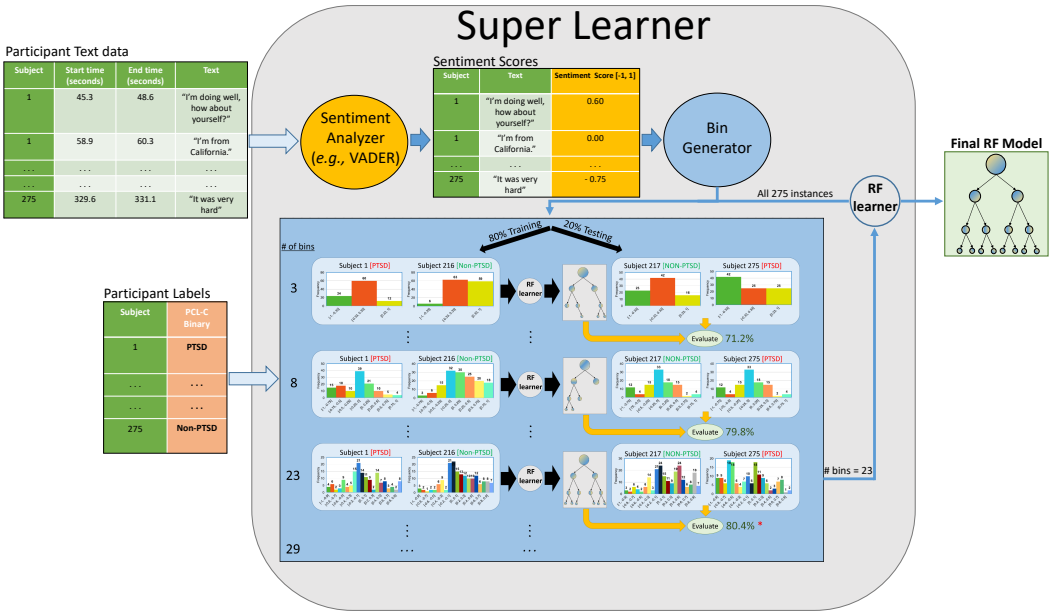



Figure 5.2: **Sentiment analysis pipeline.** A simplified version of our pipeline. Raw text is given to a sentiment analyzer (*i.e.,* VADER/Textblob/Flair) that outputs a compound scalar score between [-1, 1] for each utterance. Note that each participant provides many such utterances in the session; our SuperLearner (SL) then bins that participant's set of scores into a set of k bins. It uses internal cross-validation (on the training set) to identify the optimal number of bins and tune hyperparameters. Here, it found that 23 bins was optimal. We repeat this process with different partitioning methods with our dataset, such as 5 fold-CV and the original train-test folds. We also consider 4 different base learners, though our figure only shows the RF learner (Linear Discriminant Analysis (LDA), support vector classifier (SGD) and Random Forests (RF), and Gradient Boosting (GB)).

that put you in a good mood?' or 'What are some things that make you mad?', as well as some neutral questions [235], [246]. Ellie also provided supportive responses (*i.e.*, 'That's great' or "I'm sorry") that were used in a balanced manner throughout the interviews [246]. The questions and animated movements of Ellie were pre-recorded and designed using SmartBody, a software from USC that automates physical and verbal reactions for virtual humans [235]. After the interview was completed, they were then debriefed and given $35 as compensation for participating.

### 5.3.3 Transcription

Transcripts were transcribed and segmented by the ELAN tool from the Max Planck Institute for Psycholinguistics [251]. The transcriptions were segmented into utterances based on audio boundaries with at least 300 milliseconds of silence in the recordings. Timestamps display the length of utterances, as seen in Figure 5.2. In the current DAIC_WoZ dataset, Ellie's responses were removed from the transcribed data [246].

All transcribed interviews underwent de-identification. Human annotators scanned all utterances for mentions of names, dates, and places that could be used to narrow down an event and replaced them with special tokens [234]. Both transcriptions and de-identification were also performed by two independent annotators, and transcription discrepancies were handled by a senior annotator.

### 5.3.4 Preprocessing

Preprocessing for sentiment analysis can differ depending on the type of analyzer being used. We consider both rule-and embedding based analyzers. Some rule-based analyzers, such as VADER, handle most preprocessing steps, with an emphasis on the lexical nature of a given document. Since VADER compares the performance of its parsimonious model against human-centric baselines, the preprocessing pipeline involves limited text cleaning; see Hutto and Gilbert (2014) for the VADER preprocessing pipeline. Embedding-based analyzers, like FLAIR, require text embeddings, which represents each word as an n-dimensional vector, where similar words tend to be closer together in this n-dimensional space. Such embedding based analyzers usually involve stemming, lemmatizing, removing special characters, generating word tokens, and word embeddings [252]. For our study, we performed these preprocessing steps for FLAIR, but allowed rule-based analyzers (VADER and TextBlob) to perform these steps with their own internal functions. Regardless of the analyzer used, we employed basic text cleaning, which involved removing any brackets or quotations, digits, stop words and any upper case letters. We also removed the first 5 utterances of every participant's data, as the transcribed text involved

conversations between the experimenters and the participant shortly before the interview began.

### 5.3.5 Sentiment analysis

For each utterance from a given subject, our system produced a sentiment score based on the emotional polarity and intensity of the utterance [245]. Three different sentiment analyzers were considered for this study. VADER is a ruled-based model that uses lexical, grammatical and syntactic rules for expressing the sentiment polarity and intensity of a text. It relies on a dictionary of words that map lexical features to sentiment scores based on emotionality [245], [253]. The VADER lexicon performs better than individual human raters ($F1 = 0.96$ vs. $F1 = 0.84$) at correctly classifying sentiment of tweets into positive, neutral or negative classes (ground truth is an aggregated group mean from 20 human raters) [245]. For a given utterance, VADER generates a compound score between [-1, 1], indicating the overall sentiment and the intensity. Phrases such as 'This is horrible' will have compound scores near -1, while 'I am so happy!' will have compound scores near 1. The second sentiment analyzer was FLAIR: An easy-to-use framework for state-of-the-art NLP, a pre-trained language model that uses a mix of supervised and unsupervised techniques to capture sentiment content by examining word vectors [254]. FLAIR relies on a vector representation of words to predict sentiment annotations depending on the order of those words. Like VADER, FLAIR outputs a sentiment polarity score between [-1, 1] for a given document [254]. Lastly, TextBlob is a rule-based analyzer, that uses a pre-trained set of categorized words. Sentiments are defined based on the semantic relation and frequency of each word in a document [255]. TextBlob also outputs a polarity score between [-1, 1].

### 5.3.6 Machine learning

We generated a sentiment score for each utterance in our dataset. For each participant, we binned their sentiment scores into evenly distributed ranged bins from [-1, 1]. In predictive modeling, binning is done to transform continuous values into intervals, with the hope of optimizing the stability of predictive

performance. It can also reduce statistical noise or complexity in the variables [256], [257]. We chose this method because we wanted to normalize the length of transcripts across all participants. By binning sentiment scores, we were able to concatenate the entirety of our features into one row per subject. There is also evidence stating that certain classification models can benefit from binning numeric values [258]. For our study, we used unsupervised binning, which places variables into bins of equal range [256]. For example, if we wanted 4 bins of sentiment scores for a given subject, we would place sentiment scores from his/her utterances into these bins: [-1, -0.5, 0, 0.5, 1]. We treated the number of bins as a hyper-parameter in our ML pipeline, and let our learner select the optimal number of bins (# bins $\in [3, 6, 7, 8, 9, 12, 15, 18, 21, 23, 25, 26, 29]$) based on our evaluation metrics in the training set. The bin sizes were randomly generated between 2 and 30 to cover a large parameter space for our learner. Bin sizes below 3 and above 30 were not considered because they did not resemble a normal distribution for compound scores. Sentiment binned scores were not normalized based on the number of utterances for this pipeline. Instead, we used bin discretization to normalize the number of features. Thus, we do not account for length of transcripts between groups, as we believe that is a relevant feature for this study. As a result, we use a super learner (SL) that combines base learners (Random Forests, Gradient Boosting, Linear Discriminant Analysis, Support Vector Machine), and model configurations on the same split of data, and then uses out-of-fold predictions to select the best configurations or models [259], [260]. Our SL selects the optimal bin size (as a sort of hyperparameter), the best performing base learner, along with its hyperparameters, then runs the resulting learner (with the chosen settings), on the training dataset, to produce a final classifier. Figure 5.2 shows a simplified version of this pipeline that involves a Random Forest algorithm.

We devised many partitioning techniques on the same dataset to determine whether our models were generalizable. Originally, the DAIC_WoZ dataset used a pre-determined training, validation and test set, since it was part of an official data science competition [246]. However, for clinical purposes, we opted to use different partitioning techniques with all the data such as: Original

folds (Train-test-split) from the AVEC-19 competition, nested cross validation (5 fold-CV), and training on participants from the 2017 edition and testing on 2019 participants (2017-to-2019). Regardless of the partitioning type, our learner used internal 5 fold-CV to optimize the parameters of a model on a training fold to produce a model that is then evaluated on a held-out fold [261]. We averaged the accuracies of the test folds in our analysis. Using multiple partitioning methods helps us understand the nature of the data, and reduces the chance of a biased predefined train and test set. We wanted to be sure that this model could generalize to unseen data, hence the motivation to implement different partitioning techniques. For each model, we ran all partitioning types and compared accuracies between them.

One central issue that arises is the imbalanced class sizes between individuals with versus without PTSD. For each fold, regardless of the partitioning type, we randomly over-sampled the minority class in the training set, if the majority class had 10% more instances compared to the minority class [262]. Here, we chose minority class samples at random, with replacement using the 'imbalanced learn' package for Python (Version 0.8.0). This was done to increase the sensitivity of our model, and reduce the number of degenerate predictions from non-PTSD individuals.

Our SL considered 3 different base learners, and hyper-tuned each one with its various parameter settings. Regardless of how we partitioned the data, we only used the training data to hyper-tune the parameters. First, we used a random forest (RF) learner. An RF is an ensemble machine learning method that constructs various decision trees, first training each decision tree on a random "bagged" subset of the data. After learning this RF model, at performance time, the instance is dropped in each of the trees. These RF models have been used extensively in speech analysis for identifying PTSD and depression [263]–[265]. One study used an RF on the AVEC-17 dataset, which only included PHQ-8 scores, and achieved a precision score of 0.68 (recall = 0.65) using only text data [265]. Another study looking at depression on Twitter used an RF with VADER, and reached a precision of 0.19, but a recall of 0.96 [242]. Lastly, one study used a combination of proprietary sentiment

analyzers (labMT, LIWC and ANEW) with an RF on 243,000 tweets (produced by 63 users) of individuals with and without PTSD labels [263]. Their RF achieved an AUC of 0.89.

The second base learner is a Support Vector Classifier (SVC), which learns the separating plane that maximizes the distance to the support vectors [266]. Traditionally, SVCs have worked well with text data, perhaps due to their non-linear flexibility due to kernels. However, we only use linear kernels due to the nature of our binned data. He *et al.* 2017 used an SVC as one of their models in an n-gram based classification of individuals with PTSD versus without. The uni-gram feature set with an SVM achieved an accuracy of 80% [241]. Leiva and Freire (2017) also used an SVC with VADER. After using PCA, their SVC achieved a precision of 0.58, and a recall of 0.56 [242].

Next is Gradient Boosting (GB), which has been used in sentiment analysis for social media text, particularly related to mental health [267]–[269]. GB is a sequential boosting method that uses large trees that concentrate on misclassified observations, found by using the gradients of large residuals computed in previous iterations to refine future predictions [270]. A recent paper examined detecting anxiety based on social media data related to the COVID-19 pandemic. Their study used sentiment analysis with a number of different models, including Extreme Gradient Boosting (XGB) [269], which is a stochastic version of GB, and is computationally faster for large datasets. Their XGB model achieved an accuracy of 73.2%, but had the highest recall with 0.87 against other models such as K-nearest neighbors, SVC, RF and decision trees [269]. Another study looked at detecting anxiety and depression from social media data using word frequencies, timing, and sentiments [268]. Though their RF model achieved the best accuracy, the GB model achieved an accuracy of 79.1%, and the results were combined in an ensemble voting approach, which achieved an accuracy of 85.1%.

Our last base learner is Linear Discriminant Analysis (LDA). This method maximizes the ratio of between-class variance to within-class variance in any dataset, in an attempt to maximize separability [271]. It has been used as both a dimensionality reduction method for variables and a classification model.

LDA makes predictions by estimating the probability that unseen inputs belong to one of two distributions. The model will classify the input based on the highest probability between the two distributions [271].

The hyperparameter space is as follows: 1) SVC: 'loss' : ['hinge', 'log', 'squared_hinge', 'modified_huber'], 'alpha' : [0.0001, 0.001, 0.01, 0.1, 1, 10, 100], 'penalty' : ['L2', 'L1', 'elasticnet', 'none']. 2) GB: 'learning_rate': [0.0001, 0.001, 0.01, 0.1, 10, 100], 'n_estimators': [50, 100, 500]. 3) RF classifier: 'max_depth': [5,10, 15], 'max_features': [2, 3], 'min_samples_leaf': [2, 3, 4, 7], 'min_samples_split': [8, 10, 12], 'n_estimators': [100, 200, 300, 500]. 4) LDA: 'solver': ['svd', 'lsqr', 'eigen'], 'tol': [0.0001, 0.0002, 0.0003, 0.1, 0.01, 0.5, 0.0009, 0.09]. The classification metric used to evaluate the models were accuracy, but we also report area under curve (AUC), and f1-score. We plotted the average score from the external folds in all sentiment analyzers and all models. To summarize, we compared three different models with three different types of cross-validation, using three different sentiment analyzers.

Table 5.1: **Demographics and outcome measures.** Clinical and demographic descriptions of both groups (Non-PTSD and PTSD individuals). For testing sex differences across both groups, a chi-squared test was used. A t-test was conducted on the group means and standard deviations of both the PCL-C and the PHQ-8 scores to determine if they were statistically significant. Standard deviation can be seen within the brackets of both assessment tests. Statistical significance was set at $p$ <0.05. N/S refers to not statistically significant.

| | Non-PTSD | PTSD | Test score value | Significance |
|---|---|---|---|---|
| Sex (Male / Female) | 122 / 66 | 48 / 39 | $X_{(1,1)}$= 2.38 | N/S |
| PTSD mean score (PCL-C) | 26.54 (± 8.77) | 57.98 (± 10.70) | $t_{(273)}$= 24.06 | $p$ <0.0001 |
| Depression mean score (PHQ-8) | 4.177 (± 3.65) | 15.69 (± 3.48) | $t_{(273)}$= 23.13 | $p$ <0.0001 |

## 5.4 Results

### 5.4.1 Demographics

Table 5.1 summarizes the clinical and demographic characteristics of both groups. When examining the number of males and females in both groups, the

Table 5.2: **Demographics and outcome measures for original partitioning folds.** Clinical and demographic descriptions of both groups (Non-PTSD and PTSD individuals) in the original train-test partition from the AVEC-19 challenge. For testing sex differences across training and testing sets, a chi-squared test was used. A t-test was conducted on the group means and standard deviations of PCL-C scores to determine if they were statistically significant. Standard deviation can be seen within the brackets of both assessment tests. Statistical significance was set at $p$ <0.05. N/S refers to not statistically significant.

| | Training | | Testing | | | |
|---|---|---|---|---|---|---|
| | Non-PTSD (n = 153) | PTSD (n = 66) | Non-PTSD (n = 35) | PTSD (n = 21) | Test score value | Significance |
| Sex (Male / Female) | 93 / 60 | 34 / 32 | 29 / 6 | 14 / 7 | $X_{(2,1)} = 10.19$ | p <0.05 |
| PTSD mean score (PCL-C) | 26.23 (± 8.47) | 56.44 (± 10.29) | 27.94 (± 10.06) | 63.05 (± 10.70) | $F_{(3,\ 271)} = 227.04$ | p <0.001 |

Table 5.3: **Demographics and outcome measures for 2017-to-2019 partitioning folds.** Clinical and demographic descriptions of participants from the 2017 (training set) and 2019 competition (testing set) (Non-PTSD and PTSD individuals). For testing sex differences across training and testing sets, a chi-squared test was used. A t-test was conducted on the group means and standard deviations of PCL-C scores to determine if they were statistically significant. Standard deviation can be seen within the brackets of both assessment tests. Statistical significance was set at $p$ <0.05. N/S refers to not statistically significant.

| | Training | | Testing | | | |
|---|---|---|---|---|---|---|
| | Non-PTSD (n = 133) | PTSD (n = 56) | Non-PTSD (n = 55) | PTSD (n = 31) | Test score value | Significance |
| Sex (Male / Female) | 77 / 56 | 25 / 31 | 45 / 10 | 23 / 8 | $X_{(2,1)} = 19.1984$ | p <0.001 |
| PTSD mean score (PCL-C) | 26.42 (± 8.74) | 55.82 (± 10.66) | 26.83 (± 8.93) | 62.00 (± 9.68) | $F_{(3,\ 271)} = 230.04$ | p <0.001 |

chi-square test returned an insignificant chi-square value ($X_{(1,1)}$= 2.38, $p >$ 0.05). For the PCL-C and PHQ-8 scores, a t-test was used to determined if there was a difference between the non-PTSD and PTSD group means. There was a significance difference for the PCL-C between groups ($t_{(273)}$= 24.06, $p$ <0.0001), and there was a significant difference in the PHQ-8 test ($t_{(273)}$= 23.13, $p$ <0.0001).

Table 5.2 summarizes the clinical and demographic characteristics from the original AVEC-19 training and testing sets. When examining the number of males and females in both sets, the chi-square test returned a significant chi-square value ($X_{(2,1)}$= 10.19, $p$ <0.001). For the PCL-C scores, a one-way ANOVA was used to determined if there was a difference between the non-PTSD and PTSD group means. There was a significance difference for the PCL-C between groups ($F_{(3,271)}$= 227.04, $p$ <0.0001).

Table 5.3 summarizes the clinical and demographic characteristics from the participants included in the 2017 and 2019 versions. Here, the unique individuals from the 2017 version represented the training set, and individuals only from the 2019 version represented the testing set. When examining the number of males and females in both sets, the chi-square test returned a significant chi-square value ($X_{(2,1)}$= 19.19, $p$ <0.001). For the PCL-C scores, a one-way ANOVA was used to determined if there was a difference between the non-PTSD and PTSD group means. There was a significance difference for the PCL-C between groups ($F_{(3,271)}$= 230.04, $p$ <0.0001).

### 5.4.2 Machine Learning and Statistical Analysis

Figure 5.3 reveals the F1 scores of each model and sentiment analyzer between the various partitioning methods. In the 5-fold CV procedure seen in figure 5.3A, the RF model with VADER returned a mean accuracy of 75.6% (± 4.5% STD). The AUC was 0.72, and the F1-score was 0.83 and 0.58 for the non-PTSD and PTSD groups. The precision was 0.70, and the recall was 0.67. The optimal number of bins for this model was 18 bins based on the accuracy of the training set occurring in the grid search. Generally, the RF model outperformed all other models with all sentiment analyzers; see evaluation metrics in Table 5.4.

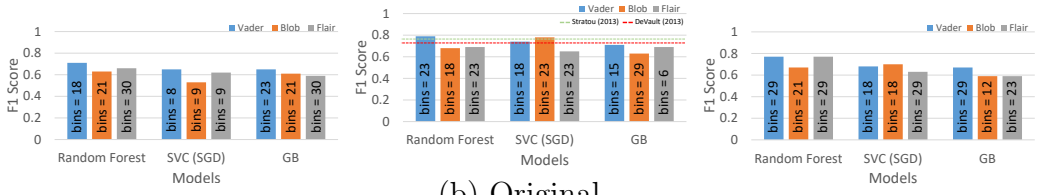
Figure 5.3B shows the results with respect to the traditional train-test-split sets from the AVEC-19 competition. The high watermark from our analysis was found here. The RF using VADER, with 23 bins, revealed the highest mean accuracy (80.4%), with an AUC of 0.80, and an F1-score of 0.85 and 0.72 for the non-PTSD and PTSD groups. The precision was 0.84 and the recall was 0.75. The RF also achieved a high accuracy with the 2017-to-2019 partitioning split (80.2%), with both the VADER and flair sentiment analyzers (bins = 23 for both). We found two benchmark studies to compare our results to. One study was from Stratou *et al.* (2013), who achieved an F1-score of 0.79 on their Naive Bayes model, with only 53 participants. However, this model incorporated audio, motioning tracking, and text data together. The second study from DeVault *et al.* (2013) achieved an accuracy of 74.4% (F1 score = 0.74) on the same 53 participants. They also used audio and textual features. The second-highest performing model was the SVC with 23 bins, using Textblob (accuracy = 78.6%), AUC = 0.78, F1-score: non-PTSD = 0.81, F1-score: PTSD = 0.75, precision = 0.77, recall = 0.75). Figure 5.3C used participants from the AVEC-2017 cohort to train, and tested on individuals only from 2019. Overall, models using VADER seemed to garner the best results compared to the other analyzers as seen in Table 5.4 (VADER= 73.2% mean, Flair = 70.3%, Textblob = 69.7%, and the RF seemed to outperform GB and SVC. Lastly, the train-test-split partitioning type had higher accuracies compared to the other partitioning methods (train-test-split mean accuracy = 74.0%, 5-CV = 69.1%, 2017-to-2019 = 70.2%).

### 5.4.3 Bin Analysis

Next, we looked at the number of sentiments across all 18 bins (the winning chosen bin size from the 5-CV partitioning method) from our SL. In Figure 5.4 and Table 5.5, the number of utterances within a certain interval of sentiment scores was plotted across both groups. For the PTSD group, there was a higher number of negative sentiments compared to the Non-PTSD group. As you move toward positive sentiments, the Non-PTSD group contained more extreme positive utterances compared to the PTSD group. A one-way ANOVA was

Table 5.4: **Performance of models across sentiment analyzers and partitioning schemes**. A list of best performing models based on the highest accuracy across the various bin sizes. The high watermark model was the RF using the VADER analyzer, with 23 bins, and the traditional train-test-split partition. This is denoted by the bolded values in the table.

| Model | Partition type | Sentiment Analyzer | Bins | Accuracy mean ± std | AUC | F1: Control | F1: Target |
|---|---|---|---|---|---|---|---|
| Random Forest | Train-test-split | **Vader** | **23** | **80.4** | **0.80** | **0.85** | **0.72** |
| | | Flair | 23 | 75.0 | 0.80 | 0.82 | 0.53 |
| | | Blob | 12 | 71.4 | 0.70 | 0.79 | 0.56 |
| | 5-Fold-CV | Vader | 18 | 75.6 ±4.5 | 0.72 | 0.83 | 0.58 |
| | | Flair | 30 | 74.0 ±2.9 | 0.70 | 0.82 | 0.49 |
| | | Blob | 21 | 70.2 ±6.3 | 0.65 | 0.79 | 0.45 |
| | 2017-to-2019 | Vader | 23 | 80.2 | 0.82 | 0.86 | 0.67 |
| | | Flair | 23 | 80.2 | 0.81 | 0.86 | 0.67 |
| | | Blob | 21 | 72.1 | 0.71 | 0.80 | 0.52 |
| Support Vector Machine (SVM) | Train-test-split | Vader | 18 | 75.0 | 0.73 | 0.80 | 0.67 |
| | | Flair | 23 | 70.0 | 0.68 | 0.78 | 0.51 |
| | | Blob | 23 | 78.6 | 0.78 | 0.81 | 0.75 |
| | 5-Fold-CV | Vader | 8 | 70.0 ±4.9 | 0.66 | 0.77 | 0.51 |
| | | Flair | 9 | 66.2 ±2.9 | 0.62 | 0.75 | 0.47 |
| | | Blob | 7 | 67.3 ±4.6 | 0.62 | 0.76 | 0.46 |
| | 2017-to-2019 | Vader | 18 | 70.1 | 0.68 | 0.77 | 0.59 |
| | | Flair | 29 | 68.6 | 0.65 | 0.78 | 0.47 |
| | | Blob | 18 | 70.0 | 0.70 | 0.74 | 0.65 |
| Linear Discriminant Analysis (LDA) | Train-test-split | Vader | 3 | 73.0 | 0.77 | 0.74 | 0.72 |
| | | Flair | 18 | 75.0 | 0.75 | 0.82 | 0.61 |
| | | Blob | 29 | 75.0 | 0.75 | 0.78 | 0.71 |
| | 5-Fold-CV | Vader | 6 | 70.2 ±2.9 | 0.67 | 0.77 | 0.58 |
| | | Flair | 30 | 64.3 ±2.5 | 0.60 | 0.73 | 0.46 |
| | | Blob | 9 | 65.5 ±6.6 | 0.61 | 0.74 | 0.48 |
| | 2017-to-2019 | Vader | 18 | 67.4 | 0.65 | 0.75 | 0.55 |
| | | Flair | 6 | 63.9 | 0.62 | 0.71 | 0.52 |
| | | Blob | 26 | 67.4 | 0.65 | 0.74 | 0.58 |
| Gradient Boosting (GB) | Train-test-split | Vader | 15 | 75.0 | 0.74 | 0.81 | 0.63 |
| | | Flair | 6 | 73.2 | 0.73 | 0.81 | 0.57 |
| | | Blob | 29 | 66.1 | 0.63 | 0.74 | 0.52 |
| | 5-Fold-CV | Vader | 23 | 70.5 ±3.2 | 0.66 | 0.79 | 0.49 |
| | | Flair | 29 | 67.3 ±2.5 | 0.60 | 0.77 | 0.41 |
| | | Blob | 21 | 67.6 ±6.2 | 0.62 | 0.77 | 0.45 |
| | 2017-to-2019 | Vader | 29 | 70.9 | 0.68 | 0.78 | 0.56 |
| | | Flair | 23 | 66.2 | 0.62 | 0.76 | 0.40 |
| | | Blob | 12 | 65.1 | 0.61 | 0.75 | 0.42 |




(a) 5-Fold CV

(b) Original Train-Test-Split

(c) 2017-to- 2019

Figure 5.3: **Machine learning results (F1 score) from partitioned types**. High watermark results from each model, sentiment analyzer and bin size for all 3 partitioning methods. The RF (23 bins) with VADER on the original train-test-split folds achieved the highest accuracy (80.4%) and an F1 score (0.79). Figure 3B displayed the benchmark F1 scores from two other studies. In the agnostic 5-fold CV, the RF (18 bins) with VADER achieved the best accuracy (75.6%, STD = ± 4.5%, F1 score = 0.71). The dashed lines represent the results from Stratou *et al.* (2013) and DeVault *et al.* (2013), who tried to predict PTSD on a smaller version of this current dataset.

conducted to examine differences between both groups across all bin intervals. A Benjamini-Hochberg correction was used to reduce type one errors across 18 comparisons. The adjusted p-value was set to 0.00284. Four intervals containing negative sentiments reflected a significant difference between both groups, as seen in Figure 5.4 and Table 5.5 ([-0.888, -0.777], [-0.555, -0.444], [-0.444, -0.333], [-0.111, 0.000]). While most of the positive sentiment bins did not return a significant p-value, the mean values for the Non-PTSD group became increasingly larger than the PTSD group.

## 5.5 Discussion

In this study, we sought to distinguish between individuals suffering from PTSD-like symptoms by analyzing emotionality during semi-structured interviews. These interviews were part of a multi-modal dataset, which was used in several data science competitions known as the AVEC challenge. We used a superlearner (SL), which combined different sentiment analyzers, features, and models to best differentiate between PTSD and non-PTSD individuals. We implemented different partitioning methods to determine whether our models could generalize well to unseen data. Our feature engineering method involved binning sentiment scores for each utterance, for each individual, and then

Table 5.5: **Average binned sentiments per group (bins = 18).** This shows the average number of sentiments for each bin per group. We conducted a one-way ANOVA to determine statistical differences between all 18 bins. and used a Benjamini-Hochberg correction to reduce type one errors. The corrected statistical significance was set to $p$ <0.00284. Statistical significance denoted by * $p$ <0.00284, ** $p$ <0.0001.

| Bin intervals | | Non-PTSD | PTSD | P-value (Adjusted B-H) |
|---|---|---|---|---|
| Start | End | Mean ± std | Mean ± std | |
| -1.000 | -0.888 | 0.266 ±0.587 | 0.471 ±0.828 | 0.18 |
| -0.888 | -0.777 | 0.734 ±0.895 | 1.057 ±1.178 | 0.145 |
| -0.777 | -0.666 | 0.814 ±0.974 | 1.563 ±1.499 | **<0.0001**** |
| -0.666 | -0.555 | 1.473 ±1.435 | 1.897 ±1.583 | 0.234 |
| -0.555 | -0.444 | 1.681 ±1.146 | 2.667 ±2.164 | **<0.0001**** |
| -0.444 | -0.333 | 2.633 ±2.233 | 4.425 ±2.970 | **<0.0001**** |
| -0.333 | -0.222 | 4.255 ±2.324 | 5.000 ±2.512 | 0.173 |
| -0.222 | -0.111 | 1.362 ±1.494 | 1.632 ±1.399 | 0.641 |
| -0.111 | 0.000 | 32.287 ±15.669 | 43.966 ±19.236 | **<0.0001**** |
| 0.000 | 0.111 | 1.213 ±1.125 | 1.827 ±1.341 | **0.003*** |
| 0.111 | 0.222 | 1.803 ±1.417 | 2.000 ±1.742 | 0.693 |
| 0.222 | 0.333 | 7.351 ±4.176 | 8.023 ±3.971 | 0.693 |
| 0.333 | 0.444 | 8.638 ±3.986 | 10.506 ±4.503 | **0.008*** |
| 0.444 | 0.555 | 4.218 ±2.522 | 4.483 ±2.832 | 0.693 |
| 0.555 | 0.666 | 4.968 ±3.103 | 5.667 ±3.194 | 0.474 |
| 0.666 | 0.777 | 4.160 ±2.477 | 3.920 ±2.161 | 0.693 |
| 0.777 | 0.888 | 4.261 ±2.808 | 3.805 ±3.013 | 0.693 |
| 0.888 | 1.000 | 3.441 ±3.696 | 2.529 ±3.066 | 0.317 |

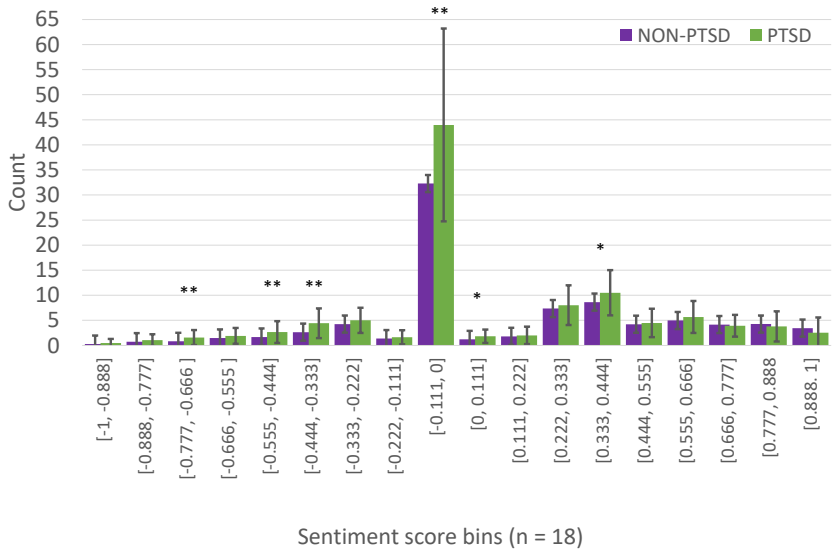



Figure 5.4: **Binned sentiment scores per group**. These are the mean values of chosen binned sentiment scores (from the RF model in the 5-fold CV partition), in each bin, for both groups. The x-axis represents the bins and their sizes, and the y-axis represents the average number of utterances that fall within those sentiment scores for both PTSD and non-PTSD groups. Table 5.5 shows the means and standard deviations, and a one-way ANOVA was conducted to compare these values in each bin between the two groups. A Benjamini-Hochberg correction was made for 18 comparisons. The adjusted p-value threshold was set to $p$ <0.00284. Error bars represent standard deviation. Significant difference denoted by * $p$ <0.00284, ** $p$ <0.0001.

concatenating them for a given subject. The SL selected the Random Forest model (RF), using the VADER analyzer, and a bin size of 23, achieving an accuracy of 80.4% (AUC = 0.80, F1 = 0.79) on the original train-test-split folds given by the AVEC-19 organizers. The partitioning method involving a training set of only 2017 AVEC participants and a test set of 2019 participants also garnered strong results. The RF model with both VADER and Flair achieved accuracies of 80.2%, and an AUC of 0.82 and 0.81. In the 5-fold CV partition, the best performing model was also the RF with VADER, but with a bin size of 18 (75.6% accuracy, ± 4.5% STD).

In the 5-fold partitioning type, our SL selected the optimal bin size of 18 as part of the final RF model. We examined the distribution of sentiments from the VADER analyzer for both groups in this bin size. We chose to analyze the bin size for the 5-fold CV partitioning, because it served as our most agnostic

validation method. Figure 5.4 revealed that individuals suffering from PTSD had more utterances in the negative sentiment bins compared to the non-PTSD group. However, this trend was slowly reversed as we moved towards the more positive sentiment bins, where non-PTSD individuals had more extremely positive sentiment scores. Our ANOVA test showed significant differences in multiple negative sentiment bins, but only one in the positive sentiment bins. These results can be seen in Table 5.5. To summarize, individuals with PTSD seemed to use negative words more often, and they used neutral words more often as well. The only words that controls used more often were extremely positive.

The results of our binned sentiment analysis are both contradictory and in concordance with several articles regarding emotional arousal in PTSD. On one hand, trauma can induce intense emotions such as anger, fear, sadness and shame, which can be reflected through language. On the other hand, individuals faced with intrusive recollections of trauma may avoid emotional eruptions, which could lead to numbness, or they may use substances to circumvent unwanted emotions. Emotional numbing is a biological process where emotions are detached from thoughts, behaviors and memories [272], [273]. Sometimes, a trauma victim may alternate between these two responses like an oscillating rhythm between an overwhelming emotional surcharge and arid states of no feeling at all [274], [275]. Based on our results, it seems some individuals with PTSD used negative sentiments and neutral sentiments more often than non-PTSD individuals. It is possible that both emotional numbing and arousal phases were present and ubiquitous in our sample size. Besides, there is large heterogeneity in behavioral responses to trauma, which may lead to high variance in detecting PTSD through language. Several studies examining the content of traumatic narratives have shown that individuals with PTSD use more emotional words, pronouns, and adjectives [276]–[278]. Pennebaker *et al.* (2003) posited that individuals suffering from suicidal ideation tend to use more emotional language and singular pronouns. The authors suggested that the increase in singular pronouns were reflective of a weakness in communicating with others [276]. Another study examining online personal

journals surrounding the 9/11 terrorist attacks showed that individuals directly affected by the attacks used stronger negative words, more first-person plural words, and less first-person singular words [279]. More recently, a longitudinal study examining 124 9/11 responders sought to predict symptom severity using an interview of their oral history. Cross-sectionally, they found greater negative language and first-person singular usage associated with symptom severity. Longitudinally, they found that anxious language was correlated with higher PCL scores, whereas first-person plural usage and longer words predicted improvement over time [280]. The novel finding in this study illustrates that language can be used to predict present and future symptom severity of PTSD. A future goal of ours is to predict symptom severity of individuals with PTSD 6 or 12 months from date of interviews. Regardless, We examined the use of pronouns in our corpus, and found that individuals without PTSD used pronouns more on average. However, individuals with PTSD used a higher proportion of pronouns compared to the rest of the tagged words in their corpus (15% vs. 14%). This is counter to the literature, however we posit that the presence of a virtual avatar may have affected the use of pronouns from individuals with PTSD.

The reason why individuals with PTSD tend to use more negative language is subject to debate. Some theorize that confronting and speaking about unpleasant emotions may help individuals cope with their trauma [281]. By reappraising their traumatic experiences, speaking with negative emotions has been shown to mediate autonomic processes that can foster and improve mental wellness [281], [282]. This has been studied in exposure therapy (ET), where individuals are repeatedly exposed to anxiety provoking stimuli until their fear response is diminished. During ET, repeated used of emotional language can allow survivors to express their emotions without experiencing the repetitive physiological sensations that come with those emotions over time [283]. One study examining emotional narratives of child abuse victims showed an association between strong negative emotional words and the depth of experiencing (meaning exploration), and emotional processing. The authors suggested that the use of strong emotional words is reflective of deeper self-

exploration and emotional processing, which can serve as catalysts to construct more positive meaning. Though we do not know the extent, nor the type of therapy some are receiving in our sample, the use of more negative language may stem from a therapeutic mechanism by which emotional language may desensitize physiological symptoms of PTSD [282], [283]. Though it is difficult to explain why more negative words were used in the PTSD group, it is possible that some have been taught to use more emotional tones to express trauma-related feelings, which is an important step for recovery.

Determining the type of emotional dysregulation early in the diagnostic procedure may benefit in tailoring specific treatments to help regulate affective responses. As mentioned, there is large heterogeneity in behavioral responses following traumatic events. Some survivors display a high degree of emotional resilience, while others go on to develop PTSD. Therefore, it is important for a clinician to reliably determine who may need therapeutic intervention, and allocate resources to those who need it most [284]. Natural language and sentiment analysis can serve as markers to facilitate initial assessments, and can be used to tailor future treatments such as mindfulness, pharmacotherapy, cognitive structuring and trauma-specific desensitization techniques such as exposure therapy or eye movement desensitization reprocessing (EMDR) [272], [285], [286]. Additionally, text-based analysis may be used to predict symptom severity several months from initial assessment [284]. Linguistic markers such as emotional language can be a non-invasive, cost-effective, rapid modality that will complement the current trend into telepsychiatry. Here, we have shown that linguistic markers such as sentiment analysis can be used to identify individuals suffering from PTSD [287].

The present study illustrates how ML techniques can be used to produce models that can identify individuals suffering from PTSD using teleconferencing interviews. Implementation and practicality of such models have become more accurate over the last 10 years. One meta-analysis examining classification studies showed that 41 of 49 articles achieved an accuracy of at least 83.7% when it came to predicting which individuals suffer from PTSD [225]. Granted, these studies used different modalities such as functional neuroimaging, fa-

cial/motion tracking and speech signals. The authors also note that learned models outperform many standard methods due to their sensitivity for hidden interactions and latent variables between predictors. They can also account for non-linear patterns in a dataset [225]. Beyond that, computational power has allowed for models to handle large, heterogeneous sources such as audio/video recordings, biological samples and neuroimaging. With enough data, these models will increase their predictive power, which may help identify certain mental conditions or even suggest certain therapies.

This study has several limitations. Firstly, the sample size is only 275 participants, with only 23% having PTSD. In our training sets, we used up-sampling to handle imbalanced classes. Ideally, a much larger dataset with more PTSD individuals could help produce a model with more robust predictions. Secondly, though the PCL-C is a reliable self-report measure, it is not considered the gold standard for diagnosing PTSD. The CAPS-5 is a clinically structured interview that is globally used to detect presence of PTSD. This would have been the preferred outcome measure for this study. Thirdly, we did not examine the presence of an artificial avatar (Ellie) conducting the interview. Though speculating, we believe that responding to an avatar may change the emotional dynamic by which an individual uses to communicate. An interesting study would look at the differences in emotional language between a human therapist and an avatar. Fourth, while we showed our model is generalizable by examining multiple partitioning methods within the AVEC-19 dataset, we do not have an external dataset (from another location) to test our model on. It is difficult to say whether our model is generalizable without testing it on a different set of participants. Lastly, a deeper analysis on part-of-speech tagging (POS) should have to be conducted in future studies. We found several studies citing an association with POS categories and severity of PTSD symptoms. POS tagging is a type of NLP which identifies the category of spoken words in a text. By categories, we refer to nouns, pronouns, adverbs and so on. A future study should consider POS tagging features to predict PTSD instead of only emotional language.

Overall, our study showed that extracting sentiment from natural language

is sufficient to detect individuals suffering from PTSD. Our analysis can be used solely, or be extended to include additional modes of data such as speech signals or motion tracking. We intend on doing this in future studies. As mentioned previously, quantifying emotional expression, valence and arousal for the purpose of diagnosis or symptom monitoring has had several issues: 1) Self-report measures rely on retrospective client or clinical insight, and do not capture emotional changes across several sessions. 2) Clinician ratings can be more objective (K <0.7), but require time intensive input and coding [288]–[290]. However, recent advances in machine learning and telepsychiatry may provide an opportunity for sentiment analysis to be implemented into online assessments and therapeutic sessions. Using ML models to detect emotion in assessments can assist in determining a diagnosis, it can be used to establish a therapeutic alliance, and it can reflect behavioral changes over time [288], [291]–[293]. Linguistic tools such as this may be used in conjunction with self-reports and interviews to better detect PTSD. With the current global pandemic, PTSD rates may continue to rise, thus accessible and accurate assessments of PTSD may be needed.

# Chapter 6

# Conclusion

While not perfect, these three projects illustrate the potential that ML has within the field of psychiatry. Specifically, we used ML to better understand the epidemiological nature of these illnesses by classifying individuals based on certain times of biological data (cognition, neural activity, and language). First, we sought to train an ML model that could detect early cognitive signs of FE-BD patients. When it comes to cognitive deficits, FE-BD individuals present closer to control individuals than they do CHR-BD individuals. Several limitations did affect this study. For example, the sample size utilized is relatively small, and while the model prediction is validated on a separate cohort of patients, the outcome should be approached with caution. Moreover, the sample sizes used in the mood state analysis are insufficient to draw any definitive conclusions. Although accuracies for specific mood states remained high, mixed and non-specific mood states exhibited lower accuracies, indicating the need for more patients with varying mood states to establish whether mood state is a confounding factor in predicting cognitive markers. However, our study successfully distinguished first-episode (FE) patients from healthy controls (HC) using cognitive test scores from the CANTAB neurocognitive battery, achieving a 76% accuracy. The model was trained on data from patients with clinically high risk (CHR) of developing bipolar disorder (BD) and HC, achieving a 77% accuracy in distinguishing between the two groups. These results suggest that cognitive impairments exist after only one manic episode for FE patients. The features from our model reflected the same cognitive

deficits identified by other studies, which could help identify FE patients at an individual level, and may aid psychiatrists in accurately diagnosing BD. Our tool represents an initial foray into transfer learning within psychiatry, demonstrating the utility of leveraging later stages of a disorder to identify early indicators and cognitive deficits. The present investigation offers a promising starting point in employing cognitive markers to authenticate the diagnostic classification of bipolar disorder (BD), particularly in its early phases. It is our aspiration that forthcoming computational models will continue to uncover and corroborate cognitive distinctions among diverse psychiatric disorders.

In the second paper, we used a data-driven approach to detect various anxiety disorders in children, as they viewed emotional facial stimuli. Using functional neuroimaging, we were able to train a superlearner that could determine which brain region could best distinguish between anxious and non-anxious children. Using ML, we discovered a novel region that has not yet been implicated in pediatric anxiety, the temporal pole. With given labels, our model was able to make individualized predictions about which children had dysregulated activity in the temporal pole. Accurate diagnosis in children is difficult, as several latent factors can offset the true underlying condition. This project has some limitations, that could be addressed in future studies. Firstly, the scientific evidence for the right temporal pole is lacking in children compared to adults. A potential future direction for this study is to conduct a meta-analysis of other studies examining anxious children or adults, since there are limited investigations focusing on the temporal pole as a region associated with anxiety. Validation of the model on adult cases could also yield interesting findings, if data is available. Another potential consideration is to test the model on a separate cohort dataset with a more homogeneous target group to improve the accuracy of the approach. The current study included children with three different types of anxiety disorders, which may have affected the overall performance accuracy. Comparing different anxiety disorders is beyond the scope of this study. Additionally, functional alignment may result in high accuracies in brain areas other than the temporal pole for the secondary analysis. Transfer learning, which involves training with one

type of labeled data and applying that model on other classes, could also be employed to examine whether the model can correctly distinguish cases based on the prior knowledge of only one type of anxiety disorder. Another potential limitation is the absence of other facial stimuli in the task. Extending the study to other related datasets with different types of visual/facial stimuli may provide further insights into how emotion is processed in the brain of anxious and non-anxious children.

Despite the potential of fMRI as a tool for understanding the neural basis of psychiatric and mental health disorders, there are several limitations that make it impractical for routine clinical use. Firstly, the high cost of fMRI equipment and the need for specialized training for both operators and interpreters limit its availability and widespread use. Moreover, the use of fMRI in a clinical setting is constrained by the need for strict experimental control, which may not be feasible in a clinical population with diverse symptoms and comorbidities. Further, fMRI measures are indirect and based on blood flow changes rather than direct measures of neural activity, which may result in ambiguous and potentially misleading interpretations. The complex and variable nature of psychiatric and mental health disorders also means that fMRI results may be difficult to interpret and generalize across patients and contexts. These limitations suggest that while fMRI holds promise as a tool for understanding the neural basis of psychiatric and mental health disorders, it is currently not a practical or reliable tool for routine clinical use.

To summarize, the study aimed to classify anxious versus non-anxious children using emotional facial stimuli and used a super learner to select Talairach regions that could best distinguish the two groups. The model achieved an accuracy above 81% for this task. The study also found that fear and angry faces can be distinguished in the temporal pole, but only after functional alignment was applied to brain scans of all subjects. The study highlights the potential of task-based fMRI designs to predict disease states and stimulus conditions, and indicates that the temporal pole is a region that requires further investigation in pediatric anxiety. This paper has significantly contributed to the existing literature on childhood anxiety and its underlying

neural mechanisms. The identification of a novel region that could potentially be associated with the pathology of pediatric anxiety is a noteworthy finding that warrants further investigation. The study's outcomes indicate that utilizing machine learning approaches to analyze face-processing, task-related functional magnetic resonance imaging (fMRI) data may facilitate the differentiation of anxious versus non-anxious children. Such an approach could potentially enhance our comprehension of the neural substrates that underlie pediatric anxiety and contribute to the validation of diagnostic categories employed by mental health professionals.

Lastly, we used natural language processing to detect individuals suffering from PTSD. Using only text data (sentiment analysis), we were able to train a model that could distinguish individuals with PTSD with over 80% accuracy. This paper identifies several limitations associated with a study aimed at using sentiment analysis to detect individuals suffering from post-traumatic stress disorder (PTSD). One of the main limitations of the study is the small sample size of only 275 participants, with only a minority having PTSD (23%). To address the issue of imbalanced classes, up-sampling was utilized during the training sets. However, it is suggested that a much larger dataset with a higher number of individuals with PTSD would be preferred for producing a model with more robust predictions. Another limitation of the study concerns the PTSD diagnostic measure used, which is the PTSD Checklist for DSM-5 (PCL-5). While this is a reliable self-report measure, it is not considered the gold standard for diagnosing PTSD. The Clinician-Administered PTSD Scale for DSM-5 (CAPS-5) is a clinically structured interview that is globally used to detect the presence of PTSD, and would have been the preferred outcome measure for this study. A third limitation pertains to the use of an artificial avatar, Ellie, to conduct the interview. Although the study did not examine this, it is suggested that the emotional dynamic by which an individual communicates may be affected when responding to an avatar. Therefore, an interesting future study would look at the differences in emotional language between a human therapist and an avatar. Furthermore, the study demonstrated the model's generalizability by examining multiple partitioning methods within the AVEC-

19 dataset. However, it is challenging to determine whether the model is generalizable without testing it on a different set of participants from another location.

Overall, the study demonstrates that sentiment analysis can be used to detect individuals suffering from PTSD. The analysis can be used alone or be extended to include additional modes of data such as speech signals or motion tracking. Recent advances in machine learning and telepsychiatry may provide an opportunity for sentiment analysis to be implemented into online assessments and therapeutic sessions. Using machine learning models to detect emotion in assessments can assist in determining a diagnosis, establish a therapeutic alliance, and reflect behavioral changes over time. Incorporating linguistic tools like sentiment analysis in conjunction with self-reports and interviews may improve the detection of PTSD, particularly in the current global pandemic, where PTSD rates may continue to rise.

This study illustrates how ML techniques can be used to identify individuals suffering from PTSD using teleconferencing interviews, something that the current pandemic has brought to light. Online clinical assessments may require additional screening measures, and ML can serve to assist in diagnostics.

In concluding, the formation of this thesis encompasses the utility of predictive analytics in psychiatry. My goal was to illustrate how machine learning can positively influence and transform translational psychiatry to focus more on individualized cases, rather than group differences. We illustrate that these models can learn information from multivariate data to predict disease outcomes on individuals rather than groups. Though my work is not the only source to showcase ML in mental health, it is part of a growing field that seeks to fully utilize the vasts amounts of human data that technology has allowed for in the 21st century. In my conceptual review, I discussed the major barriers that precision health needs to overcome before such a paradigm can be implemented. Barriers such as data ethics and privacy, availability and representativeness, noisy diagnostic labels and the role of AI in precision medicine are significant obstacles that need to be addressed, should computational psychiatry forge ahead. But currently, several research initiatives are proving why ML can

ultimately serve to help individuals understand their own condition, and recover faster.

The above papers demonstrate the application of ML for nuanced problems within the field of psychiatry. Namely, we focused on using ML in a clinical setting, where current diagnostics are unclear, and assessment techniques are complex, time-consuming and costly [8]. Instead of group-based statistical analysis and univariate testing, ML can be leveraged to detect multivariate, non-linear patterns that can be found at the individual level. The multivariate patterns can relate to biological, social, or psychological factors, all of which can be examined by ML algorithms in some form. ML can also provide an alternative examination of underlying mechanisms that exist in these illnesses. The introduction of the RDOC should propel us further into these systems, allowing us an alternative view of these diseases besides symptomatology. With such tools, we can specify where to look, and what to look for when it comes to these illnesses. However, in the conceptual review, we presented possible barriers that stand between an optimal conjoining of ML and psychiatry. There is still a long road ahead, but these new tools can ultimately serve us to learn more about psychiatric illnesses than ever before.

# References


[1] J. L. J. Quah, S. Yap, S. O. Cheah, *et al.*, "Knowledge of signs and symptoms of heart attack and stroke among singapore residents," *BioMed research international*, vol. 2014, 2014.

[2] B. I. Scheel and K. Holtedahl, "Symptoms, signs, and tests: The general practitioner's comprehensive approach towards a cancer diagnosis," *Scandinavian journal of primary health care*, vol. 33, no. 3, pp. 170–177, 2015.

[3] A. Azzarelli, A. Guzzon, S. Pilotti, V. Quagliuolo, A. Bono, and S. Di Pietro, "Accuracy of breast cancer diagnosis by physical, radiologic and cytologic combined examinations," *Tumori Journal*, vol. 69, no. 2, pp. 137–141, 1983.

[4] T. C. Manschreck and A. M. Kleinman, "Psychiatry's identity crisis: A critical rational remedy," *General hospital psychiatry*, vol. 1, no. 2, pp. 166–173, 1979.

[5] A. Surıs, R. Holliday, and C. S. North, "The evolution of the classification of psychiatric disorders," *Behavioral Sciences*, vol. 6, no. 1, p. 5, 2016.

[6] A. Frances, M. B. First, and H. A. Pincus, *DSM-IV guidebook*. American Psychiatric Association, 1995.

[7] R. Freedman, D. A. Lewis, R. Michels, *et al.*, *The initial field trials of dsm-5: New blooms and old thorns*, 2013.

[8] D. B. Dwyer, P. Falkai, and N. Koutsouleris, "Machine learning approaches for clinical psychology and psychiatry," *Annual review of clinical psychology*, vol. 14, pp. 91–118, 2018.

[9] L. Wunderink, S. Sytema, F. J. Nienhuis, and D. Wiersma, "Clinical recovery in first-episode psychosis," *Schizophrenia Bulletin*, vol. 35, no. 2, pp. 362–369, 2009.

[10] J. Hegarty, R. Baldessarini, M. Tohen, C. Waternaux, and G. Oepen, *One hundred years of schizophrenia: A meta-analysis of the outcome literature american journal of psychiatry, 15*, 1994.

[11] S. G. Hofmann, A. Asnaani, I. J. Vonk, A. T. Sawyer, and A. Fang, "The efficacy of cognitive behavioral therapy: A review of meta-analyses," *Cognitive therapy and research*, vol. 36, no. 5, pp. 427–440, 2012.

[12] A. J. Rush, M. H. Trivedi, S. R. Wisniewski, *et al.*, "Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A star* d report," *American Journal of Psychiatry*, vol. 163, no. 11, pp. 1905–1917, 2006.

[13] E. H. Wong, F. Yocca, M. A. Smith, and C.-M. Lee, "Challenges and opportunities for drug discovery in psychiatric disorders: The drug hunters' perspective," *The The International Journal of Neuropsychopharmacology*, vol. 13, no. 9, pp. 1269–1284, 2010.

[14] S. E. Hyman, "The diagnosis of mental disorders: The problem of reification," *Annual review of clinical psychology*, vol. 6, pp. 155–179, 2010.

[15] E. Ramsden, "From rodent utopia to urban hell: Population, pathology, and the crowded rats of nimh," *Isis*, vol. 102, no. 4, pp. 659–688, 2011.

[16] T. R. Insel, "The nimh research domain criteria (rdoc) project: Precision medicine for psychiatry," *American Journal of Psychiatry*, vol. 171, no. 4, pp. 395–397, 2014.

[17] T. Insel. "Toward a new understanding of mental illness," Ted Talks. (2013), [Online]. Available: `https://www.youtube.com/watch?v=PeZ-U0pj9LI`.

[18] World Health Organization, *Suicide fact sheet*, `https://www.who.int/news-room/fact-sheets/detail/suicide`, Last accessed on 2022-01-12, 2021.

[19] ——, *Depression fact sheet*, `https://www.who.int/news-room/fact-sheets/detail/depression`, Last accessed on 2022-01-12, 2021.

[20] C. for Disease Control, Prevention, *et al.*, "Data and statistics on children's mental health," *Retrieved April*, vol. 17, p. 2019, 2019.

[21] A. H. Mokdad, K. Ballestros, M. Echko, *et al.*, "The state of us health, 1990-2016: Burden of diseases, injuries, and risk factors among us states," *Jama*, vol. 319, no. 14, pp. 1444–1472, 2018.

[22] H. A. Whiteford, A. J. Ferrari, L. Degenhardt, V. Feigin, and T. Vos, "The global burden of mental, neurological and substance use disorders: An analysis from the global burden of disease study 2010," *PloS one*, vol. 10, no. 2, e0116820, 2015.

[23] K.-L. Lim, P. Jacobs, A. Ohinmaa, D. Schopflocher, and C. S. Dewa, "A new population-based measure of the economic burden of mental illness in canada," *Chronic Dis Can*, vol. 28, no. 3, pp. 92–98, 2008.

[24] C. de Oliveira, J. Cheng, S. Vigod, J. Rehm, and P. Kurdyak, "Patients with high mental health costs incur over 30 percent more costs than other high-cost patients," *Health Affairs*, vol. 35, no. 1, pp. 36–43, 2016.

[25] M. H. C. of Canada, "Making the case for investing in mental health in canada," *London (ON): Mental Health Commission*, 2013.

[26] S. O. Lilienfeld and M. T. Treadway, "Clashing diagnostic approaches: Dsm-icd versus rdoc," *Annual review of clinical psychology*, vol. 12, pp. 435–463, 2016.

[27] B. J. Carroll, "The dexamethasone suppression test for melancholia," *The British journal of psychiatry*, vol. 140, no. 3, pp. 292–304, 1982.

[28] D. J. Kupfer, F. G. Foster, P. Coble, R. J. McPartland, and R. F. Ulrich, "The application of eeg sleep for the differential diagnosis of affective disorders.," *The American journal of psychiatry*, 1978.

[29] W. G. Iacono, V. B. Tuason, and R. A. Johnson, "Dissociation of smooth-pursuit and saccadic eye tracking in remitted schizophrenics: An ocular reaction time task that schizophrenics perform well," *Archives of General Psychiatry*, vol. 38, no. 9, pp. 991–996, 1981.

[30] D. Bzdok and B. T. Yeo, "Inference in the age of big data: Future perspectives on neuroscience," *Neuroimage*, vol. 155, pp. 549–564, 2017.

[31] C. G. Begley and L. M. Ellis, "Raise standards for preclinical cancer research," *Nature*, vol. 483, no. 7391, pp. 531–533, 2012.

[32] A. Eklund, M. Andersson, C. Josephson, M. Johannesson, and H. Knutsson, "Does parametric fmri analysis with spm yield valid results?—an empirical study of 1484 rest datasets," *NeuroImage*, vol. 61, no. 3, pp. 565–578, 2012.

[33] A. Eklund, T. E. Nichols, and H. Knutsson, "Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates," *Proceedings of the national academy of sciences*, vol. 113, no. 28, pp. 7900–7905, 2016.

[34] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain.," *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[35] K. Rost, G. R. Smith, D. B. Matthews, and B. Guise, "The deliberate misdiagnosis of major depression in primary care," *Archives of family medicine*, vol. 3, no. 4, p. 333, 1994.

[36] I. C. Passos, P. Ballester, F. D. Rabelo-da-Ponte, and F. Kapczinski, "Precision psychiatry: The future is now," *The Canadian Journal of Psychiatry*, p. 0 706 743 721 998 044, 2021.

[37] J. Sawalha, L. Cao, J. Chen, *et al.*, "Individualized identification of first-episode bipolar disorder using machine learning and cognitive tests," *Journal of Affective Disorders*, vol. 282, pp. 662–668, 2021.

[38] D. C. Baumgart, "Digital advantage in the covid-19 response: Perspective from canada's largest integrated digitalized healthcare system," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–4, 2020.

[39] I. C. Passos, B. Mwangi, and F. Kapczinski, *Personalized psychiatry: Big data analytics in mental health*. Springer, 2019.

[40] C. Jones, L. Smith-MacDonald, A. Miguel-Cruz, *et al.*, "Virtual reality–based treatment for military members and veterans with combat-related posttraumatic stress disorder: Protocol for a multimodular motion-assisted memory desensitization and reconsolidation randomized controlled trial," *JMIR research protocols*, vol. 9, no. 10, e20620, 2020.

[41] K. M. Kostick-Quenet, I. G. Cohen, S. Gerke, *et al.*, "Mitigating racial bias in machine learning," *Journal of Law, Medicine & Ethics*, vol. 50, no. 1, pp. 92–100, 2022.

[42] Y. Merino, L. Adams, and W. J. Hall, *Implicit bias and mental health professionals: Priorities and directions for research*, 2018.

[43] R. Y. Coley, E. Johnson, G. E. Simon, M. Cruz, and S. M. Shortreed, "Racial/ethnic disparities in the performance of prediction models for death by suicide after mental health visits," *JAMA psychiatry*, vol. 78, no. 7, pp. 726–734, 2021.

[44] J. Bulao, "How much data is created every day in 2021," *techjury*, 2021.

[45] M. De Choudhury and S. De, "Mental health discourse on reddit: Self-disclosure, social support, and anonymity," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[46] J. Berg, "The e-health revolution and the necessary evolution of informed consent," *Ind. Health L. Rev.*, vol. 11, p. 589, 2014.

[47] N. C. Jacobson, K. H. Bentley, A. Walton, *et al.*, "Ethical dilemmas posed by mobile health and machine learning in psychiatry research," *Bulletin of the World Health Organization*, vol. 98, no. 4, p. 270, 2020.

[48] N. Martinez-Martin, K. Kreitmair, *et al.*, "Ethical issues for direct-to-consumer digital psychotherapy apps: Addressing accountability, data protection, and consent," *JMIR mental health*, vol. 5, no. 2, e9423, 2018.

[49] N. A. M. H. Council, *Opportunities and challenges of developing information technologies on behavioral and social science clinical research*, 2018.

[50] J. Sawalha, M. Yousefnezhad, Z. Shah, M. R. Brown, A. J. Greenshaw, and R. Greiner, "Detecting presence of ptsd using sentiment analysis from text data," *Frontiers in Psychiatry*, p. 2618,

[51] A. Aboraya, E. Rankin, C. France, A. El-Missiry, and C. John, "The reliability of psychiatric diagnosis revisited: The clinician's guide to improve the reliability of psychiatric diagnosis," *Psychiatry (Edgmont)*, vol. 3, no. 1, p. 41, 2006.

[52] G. Starke, E. De Clercq, S. Borgwardt, and B. S. Elger, "Computing schizophrenia: Ethical challenges for machine learning in psychiatry," *Psychological Medicine*, vol. 51, no. 15, pp. 2515–2521, 2021.

[53] K. S. Kendler, "The nature of psychiatric disorders," *World Psychiatry*, vol. 15, no. 1, pp. 5–12, 2016.

[54] P. Zachar, "Psychiatric disorders: Natural kinds made by the world or practical kinds made by us?" *World Psychiatry*, vol. 14, no. 3, p. 288, 2015.

[55] J. D. Gabrieli, S. S. Ghosh, and S. Whitfield-Gabrieli, "Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience," *Neuron*, vol. 85, no. 1, pp. 11–26, 2015.

[56] T. Insel, B. Cuthbert, M. Garvey, *et al.*, *Research domain criteria (rdoc): Toward a new classification framework for research on mental disorders*, 2010.

[57] T. R. Insel and B. N. Cuthbert, "Brain disorders? precisely," *Science*, vol. 348, no. 6234, pp. 499–500, 2015.

[58] M. Brunn, A. Diefenbacher, P. Courtet, and W. Genieys, "The future is knocking: How artificial intelligence will fundamentally change psychiatry," *Academic Psychiatry*, vol. 44, no. 4, pp. 461–466, 2020.

[59] G. E. Simon and B. J. Yarborough, "Good news: Artificial intelligence in psychiatry is actually neither," *Psychiatric Services*, vol. 71, no. 3, pp. 219–220, 2020.

[60] Y.-h. Sheu, "Illuminating the black box: Interpreting deep neural network models for psychiatric research," *Frontiers in Psychiatry*, p. 1091, 2020.

[61] C. Chandler, P. W. Foltz, and B. Elvevåg, "Using machine learning in psychiatry: The need to establish a framework that nurtures trustworthiness," *Schizophrenia bulletin*, vol. 46, no. 1, pp. 11–14, 2020.

[62] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti, "A survey of methods for explaining black box models (2018)," *arXiv preprint arXiv:1802.01933*, 2018.

[63] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," in *Explainable AI: interpreting, explaining and visualizing deep learning*, Springer, 2019, pp. 5–22.

[64] A. Prahl and L. M. Van Swol, "Out with the humans, in with the machines?: Investigating the behavioral and psychological effects of replacing human advisors with a machine," *Human-Machine Communication*, vol. 2, pp. 209–234, 2021.

[65] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err.," *Journal of Experimental Psychology: General*, vol. 144, no. 1, p. 114, 2015.

[66] B. C. Kok and H. Soh, "Trust in robots: Challenges and opportunities," *Current Robotics Reports*, vol. 1, pp. 297–309, 2020.

[67] S. Booth, J. Tompkin, H. Pfister, J. Waldo, K. Gajos, and R. Nagpal, "Piggybacking robots: Human-robot overtrust in university dormitory security," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 426–434.

[68] F. Cabitza, R. Rasoini, and G. F. Gensini, "Unintended consequences of machine learning in medicine," *Jama*, vol. 318, no. 6, pp. 517–518, 2017.

[69] S. Gaube, H. Suresh, M. Raue, *et al.*, "Do as ai say: Susceptibility in deployment of clinical decision-aids," *NPJ digital medicine*, vol. 4, no. 1, p. 31, 2021.

[70] J. Anderson and L. Rainie, "The future of well-being in a tech-saturated world," *Pew Research Center*, vol. 17, 2018.

[71] K. R. Merikangas, M. Ames, L. Cui, *et al.*, "The impact of comorbidity of mental and physical conditions on role disability in the US adult household population," *Arch. Gen. Psychiatry*, vol. 64, no. 10, pp. 1180–1188, Oct. 2007.

[72] T. Moradi, P. Allebeck, A. Jacobsson, and C. Mathers, "The burden of disease in Sweden measured with DALY: Neuropsychiatric diseases and cardiovascular diseases dominate," *Lakartidningen*, vol. 103, no. 3, pp. 137–141, 2006.

[73] J. S. Manning, *Burden of Illness in Bipolar Depression*, 06. 2005, vol. 07, Publication Title: The Primary Care Companion to The Journal of Clinical Psychiatry.

[74] B. Solé, E. Jiménez, C. Torrent, *et al.*, "Cognitive Impairment in Bipolar Disorder: Treatment and Prevention Strategies," *Int. J. Neuropsychopharmacol.*, vol. 20, no. 8, pp. 670–680, Aug. 2017.

[75] M. Vrabie, V. Marinescu, A. Talaşman, O. Tăutu, E. Drima, and I. Micluţia, "Cognitive impairment in manic bipolar patients: Important, understated, significant aspects," en, *Annals of General Psychiatry*, vol. 14, no. 1, p. 41, Dec. 2015, ISSN: 1744-859X. DOI: 10.1186/s12991-015-0080-0. [Online]. Available: http://www.annals-general-psychiatry.com/content/14/1/41 (visited on 08/26/2020).

[76] B. Bortolato, K. W. Miskowiak, C. A. Köhler, E. Vieta, and A. F. Carvalho, "Cognitive dysfunction in bipolar disorder and schizophrenia: A systematic review of meta-analyses," *Neuropsychiatr. Dis. Treat.*, vol. 11, pp. 3111–3125, Dec. 2015.

[77] C. Bourne, Ö. Aydemir, V. Balanzá-Martínez, *et al.*, "Neuropsychological testing of cognitive impairment in euthymic bipolar disorder: An individual patient data meta-analysis," eng, *Acta Psychiatrica Scandinavica*, vol. 128, no. 3, pp. 149–162, Sep. 2013, ISSN: 1600-0447. DOI: 10.1111/acps.12133.

[78] L. J. Robinson, J. M. Thompson, P. Gallagher, *et al.*, "A meta-analysis of cognitive deficits in euthymic patients with bipolar disorder," *J. Affect. Disord.*, vol. 93, no. 1-3, pp. 105–115, 2006.

[79] A. Martínez-Arán, E. Vieta, M. Reinares, *et al.*, "Cognitive Function Across Manic or Hypomanic, Depressed, and Euthymic States in Bipolar Disorder," *American Journal of Psychiatry*, vol. 161, no. 2, pp. 262–270, 2004, ISBN: 0002953X (ISSN), ISSN: 0002953X. DOI: 10.1176/appi.ajp.161.2.262.

[80] J. H. MacCabe, M. P. Lambe, S. Cnattingius, *et al.*, *Excellent school performance at age 16 and risk of adult bipolar disorder: national cohort study*, 2. 2010, vol. 196, Publication Title: British Journal of Psychiatry.

[81] B. Arts, N. Jabben, L. Krabbendam, and J. van Os, "Meta-analyses of cognitive functioning in euthymic bipolar patients and their first-degree relatives," *Psychol. Med.*, vol. 38, no. 6, pp. 771–785, 2008.

[82] E. Bora, M. Yucel, and C. Pantelis, "Cognitive endophenotypes of bipolar disorder: A meta-analysis of neuropsychological deficits in euthymic patients and their first-degree relatives," *J. Affect. Disord.*, vol. 113, no. 1-2, pp. 1–20, 2009.

[83] I. J. Torres, V. G. DeFreitas, C. M. DeFreitas, *et al.*, *Neurocognitive Functioning in Patients With Bipolar I Disorder Recently Recovered From a First Manic Episode*, 09. 2010, vol. 71, Publication Title: The Journal of Clinical Psychiatry.

[84] R. Daglas, M. Yücel, S. Cotton, K. Allott, S. Hetrick, and M. Berk, "Cognitive impairment in first-episode mania: A systematic review of the evidence in the acute and remission phases of the illness," *Int J Bipolar Disord*, vol. 3, p. 9, Apr. 2015.

[85] H. H. Elshahawi, H. Essawi, M. A. Rabie, M. Mansour, Z. A. Beshry, and A. N. Mansour, "Cognitive functions among euthymic bipolar I patients after a single manic episode versus recurrent episodes," *J. Affect. Disord.*, vol. 130, no. 1-2, pp. 180–191, 2011.

[86] M.-J. Wu, I. C. Passos, I. E. Bauer, *et al.*, "Individualized identification of euthymic bipolar disorder using the Cambridge Neuropsychological Test Automated Battery (CANTAB) and machine learning," *J. Affect. Disord.*, vol. 192, pp. 219–225, Mar. 2016.

[87] A. G. Isasi, E. Echeburúa, J. M. Limiñana, and A. González-Pinto, "How effective is a psychological intervention program for patients with refractory bipolar disorder? A randomized controlled trial," *J. Affect. Disord.*, vol. 126, no. 1-2, pp. 80–87, Oct. 2010.

[88] A. R. Rosa, I. González-Ortega, A. González-Pinto, *et al.*, "One-year psychosocial functioning in patients in the early vs. late stage of bipolar disorder," eng, *Acta Psychiatrica Scandinavica*, vol. 125, no. 4, pp. 335–341, Apr. 2012, ISSN: 1600-0447. DOI: 10.1111/j.1600-0447.2011.01830.x.

[89] B. Cha, J. H. Kim, T. H. Ha, J. S. Chang, and K. Ha, "Polarity of the first episode and time to diagnosis of bipolar I disorder," *Psychiatry Investig.*, vol. 6, no. 2, pp. 96–101, 2009.

[90] B. G. Schimmelmann, P. Conus, J. Edwards, P. D. McGorry, and M. Lambert, "Diagnostic stability 18 months after treatment initiation for first-episode psychosis," *J. Clin. Psychiatry*, vol. 66, no. 10, pp. 1239–1246, Oct. 2005.

[91] E. Salagre, S. Dodd, A. Aedo, *et al.*, "Toward precision psychiatry in bipolar disorder: Staging 2.0," *Frontiers in psychiatry*, vol. 9, p. 641, 2018.

[92] J. Walsh-Messinger, H. Jiang, H. Lee, K. Rothman, H. Ahn, and D. Malaspina, "Relative importance of symptoms, cognition, and other multilevel variables for psychiatric disease classifications by machine learning," en, *Psychiatry Research*, vol. 278, pp. 27–34, Aug. 2019, ISSN: 0165-1781. DOI: 10.1016/j.psychres.2019.03.048. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165178118321899 (visited on 11/30/2020).

[93] M. Berk, K. Hallam, G. S. Malhi, *et al.*, "Evidence and implications for early intervention in bipolar disorder," *J. Ment. Health*, vol. 19, no. 2, pp. 113–126, 2010.

[94] R. S. C. Lee, D. F. Hermens, J. Scott, *et al.*, "A meta-analysis of neuropsychological functioning in first-episode bipolar disorders," eng, *Journal of Psychiatric Research*, vol. 57, pp. 1–11, Oct. 2014, ISSN: 1879-1379. DOI: 10.1016/j.jpsychires.2014.06.019.

[95] M. N. Levaux, S. Potvin, A. A. Sepehry, J. Sablier, A. Mendrek, and E. Stip, "Computerized assessment of cognition in schizophrenia: Promises and pitfalls of CANTAB," *European Psychiatry*, vol. 22, no. 2, pp. 104–115, 2007, ISSN: 09249338. DOI: 10.1016/j.eurpsy.2006.11.004.

[96] J. A. Sweeney, J. A. Kmiec, and D. J. Kupfer, "Neuropsychologic impairments in bipolar and unipolar mood disorders on the CANTAB neurocognitive battery," *Biological Psychiatry*, vol. 48, no. 7, pp. 674–684, 2000, ISSN: 00063223. DOI: 10.1016/S0006-3223(00)00910-0.

[97] M. M. Kurtz and R. T. Gerraty, "A Meta-analytic Investigation of Neurocognitive Deficitis in Bipolar Illness," vol. 23, no. 5, pp. 551–562, 2010. DOI: 10.1037/a0016277.A.

[98] D. S. Cha, N. E. Carmona, M. Subramaniapillai, *et al.*, "Cognitive impairment as measured by the THINC-integrated tool (THINC-it): Association with psychosocial function in major depressive disorder," en, *Journal of Affective Disorders*, vol. 222, pp. 14–20, Nov. 2017, ISSN: 01650327. DOI: 10.1016/j.jad.2017.06.036. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S016503271632208X (visited on 08/26/2020).

[99] R. W. Lam, S. H. Kennedy, R. S. McIntyre, and A. Khullar, "Cognitive Dysfunction in Major Depressive Disorder: Effects on Psychosocial Functioning and Implications for Treatment," en, *The Canadian Journal of Psychiatry*, vol. 59, no. 12, pp. 649–654, Dec. 2014, ISSN: 0706-7437, 1497-0015. DOI: 10.1177/070674371405901206. [Online]. Available: http://journals.sagepub.com/doi/10.1177/070674371405901206 (visited on 08/26/2020).

[100] E. E. Michalak, L. N. Yatham, S. Kolesar, and R. W. Lam, "Bipolar Disorder and Quality of Life: A Patient-Centered Perspective," en, *Quality of Life Research*, vol. 15, no. 1, pp. 25–37, Feb. 2006, ISSN: 0962-9343, 1573-2649. DOI: 10.1007/s11136-005-0376-7. [Online]. Available: http://link.springer.com/10.1007/s11136-005-0376-7 (visited on 08/26/2020).

[101] E. E. Michalak, I. J. Torres, D. J. Bond, R. W. Lam, and L. N. Yatham, "The relationship between clinical outcomes and quality of life in first-episode mania: A longitudinal analysis: Clinical outcomes and quality of life in FEM," en, *Bipolar Disorders*, vol. 15, no. 2, pp. 188–198, Mar. 2013, ISSN: 13985647. DOI: 10.1111/bdi.12049. [Online]. Available: http://doi.wiley.com/10.1111/bdi.12049 (visited on 08/26/2020).

[102] M. R. Basso, N. Lowery, J. Neel, R. Purdie, and R. A. Bornstein, "Neuropsychological impairment among manic, depressed, and mixed-episode inpatients with bipolar disorder," eng, *Neuropsychology*, vol. 16, no. 1, pp. 84–91, Jan. 2002, ISSN: 0894-4105. DOI: 10.1037//0894-4105.16.1.84.

[103] E. Bora, M. Yücel, and C. Pantelis, "Cognitive impairment in schizophrenia and affective psychoses: Implications for dsm-v criteria and beyond," *Schizophrenia Bulletin*, vol. 36, no. 1, pp. 36–42, 2010, ISSN: 05867614. DOI: 10.1093/schbul/sbp094.

[104] D. C. Glahn, P. M. Thompson, and J. Blangero, "Neuroimaging endophenotypes: Strategies for finding genes influencing brain structure and function," *Human Brain Mapping*, vol. 28, no. 6, pp. 488–501, 2007, ISSN: 10659471. DOI: 10.1002/hbm.20401.

[105] M. Sanches, I. E. Bauer, J. F. Galvez, G. B. Zunta-Soares, and J. C. Soares, "The Management of Cognitive Impairment in Bipolar Disorder," *American Journal of Therapeutics*, vol. 22, no. 6, pp. 477–

486, 2015, ISBN: 0000000000000, ISSN: 1075-2765. DOI: 10.1097/mjt.0000000000000120.

[106] D. Frydecka, A. M. Eissa, D. H. Hewedi, *et al.*, "Impairments of working memory in schizophrenia and bipolar disorder: The effect of history of psychotic symptoms and different aspects of cognitive task demands," *Frontiers in Behavioral Neuroscience*, vol. 8, no. NOV, pp. 1–11, 2014, ISSN: 16625153. DOI: 10.3389/fnbeh.2014.00416.

[107] C. Jahshan, K. J. Wynn, M. Amanda, C. D. Glahn, L. Altshuler, and M. Green, "Cross-Diagnostic Comparison of Visual Processing in Bipolar Disorder and Schizophrenia," *Journal of Psychiatric Research*, vol. 28, no. 1, pp. 7–12, 2016, ISBN: 0000000000000. DOI: 10.1097/YCO.0000000000000122.Reward.

[108] T. Maekawa, S. Katsuki, J. Kishimoto, *et al.*, "Altered visual information processing systems in bipolar disorder: Evidence from visual MMN and P3," *Frontiers in Human Neuroscience*, vol. 7, no. JUL, pp. 1–11, 2013, ISSN: 16625161. DOI: 10.3389/fnhum.2013.00403.

[109] E. Yang, D. Tadin, D. M. Glasser, S. Wook Hong, R. Blake, and S. Park, "Visual context processing in bipolar disorder: A comparison with schizophrenia," *Frontiers in Psychology*, vol. 4, no. August, pp. 1–12, 2013, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2013.00569.

[110] A. Duffy, "The early course of bipolar disorder in youth at familial risk," *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, vol. 18, no. 3, pp. 200–205, 2009, ISSN: 17198429.

[111] R. H. Perlis, C. S. Perlis, Y. Wu, C. Hwang, M. Joseph, and A. A. Nierenberg, "Industry sponsorship and financial conflict of interest in the reporting of clinical trials in psychiatry," *American Journal of Psychiatry*, vol. 162, no. 10, pp. 1957–1960, 2005.

[112] A. Serretti, "The present and future of precision medicine in psychiatry: Focus on clinical psychopharmacology of antidepressants," *Clinical Psychopharmacology and Neuroscience*, vol. 16, no. 1, p. 1, 2018.

[113] A. Ströhle, J. Gensichen, and K. Domschke, "The diagnosis and treatment of anxiety disorders," en, *Deutsches Aerzteblatt Online*, vol. 115, no. 37, p. 611, Sep. 2018, ISSN: 1866-0452. DOI: 10.3238/arztebl.2018.0611. [Online]. Available: https://www.aerzteblatt.de/10.3238/arztebl.2018.0611 (visited on 04/14/2020).

[114] K. R. Merikangas, E. F. Nakamura, and R. C. Kessler, "Epidemiology of mental disorders in children and adolescents," *Dialogues in Clinical Neuroscience*, vol. 11, no. 1, pp. 7–20, Mar. 2009, ISSN: 1294-8322. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2807642/ (visited on 03/25/2020).

[115] Kessler, R.C., Avenevoli, S., Costello, E.J., Georgiades, K., Green, J.G., Gruber, M.J., He, J.P., Koretz, D., McLaughlin, K.A., Petukhova, M. and Sampson, N.A., "Prevalence, persistence, and sociodemographic correlates of DSM-IV disorders in the National Comorbidity Survey Replication Adolescent Supplement," *Archives of general psychiatry*, vol. 69, no. 4, pp. 372–380, Apr. 2012, ISSN: 0003-990X. DOI: 10.1001/archgenpsychiatry.2011.160. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3445020/ (visited on 03/25/2020).

[116] B. Bandelow and S. Michaelis, "Epidemiology of anxiety disorders in the 21st century," *Dialogues in Clinical Neuroscience*, vol. 17, no. 3, pp. 327–335, Sep. 2015, ISSN: 1294-8322. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4610617/ (visited on 03/25/2020).

[117] A. L. Carpenter, D. B. Pincus, E. C. Perrin, M. H. Bair-Merritt, and N. D. Mian, "Early identification of anxiety disorders: The role of the pediatrician in primary care," en, *Children's Health Care*, vol. 47, no. 1, pp. 34–50, Jan. 2018, ISSN: 0273-9615, 1532-6888. DOI: 10.1080/02739615.2016.1275642. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/02739615.2016.1275642 (visited on 03/25/2020).

[118] Kessler, R.C., Berglund, P., Demler, O., Jin, R., Merikangas, K.R. and Walters, E.E., "Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication," eng, *Archives of General Psychiatry*, vol. 62, no. 6, pp. 593–602, Jun. 2005, ISSN: 0003-990X. DOI: 10.1001/archpsyc.62.6.593.

[119] R. M. Rapee, "The preventative effects of a brief, early intervention for preschool-aged children at risk for internalising: Follow-up into middle adolescence: Adolescent effects of brief, early intervention for internalising," en, *Journal of Child Psychology and Psychiatry*, vol. 54, no. 7, pp. 780–788, Jul. 2013, ISSN: 00219630. DOI: 10.1111/jcpp.12048. [Online]. Available: http://doi.wiley.com/10.1111/jcpp.12048 (visited on 03/25/2020).

[120] Reeb-Sutherland, B.C., Rankin Williams, L., Degnan, K.A., Pérez-Edgar, K., Chronis-Tuscano, A., Leibenluft, E., Pine, D.S., Pollak, S.D. and Fox, N.A., "Identification of emotional facial expressions among behaviorally inhibited adolescents with lifetime anxiety disorders," en, *Cognition and Emotion*, vol. 29, no. 2, pp. 372–382, Feb. 2015, ISSN: 0269-9931, 1464-0600. DOI: 10.1080/02699931.2014.913552. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/02699931.2014.913552 (visited on 08/05/2020).

[121] Y. Bar-Haim, D. Lamy, L. Pergamin, M. J. Bakermans-Kranenburg, and M. H. van Ijzendoorn, "Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study," eng, *Psychological*

*Bulletin*, vol. 133, no. 1, pp. 1–24, Jan. 2007, ISSN: 0033-2909. DOI: 10.1037/0033-2909.133.1.1.

[122] I. R. Olson, A. Plotzker, and Y. Ezzyat, "The enigmatic temporal pole: A review of findings on social and emotional processing," en, *Brain*, vol. 130, no. 7, pp. 1718–1731, May 2007, ISSN: 0006-8950, 1460-2156. DOI: 10.1093/brain/awm052. [Online]. Available: https://academic.oup.com/brain/article-lookup/doi/10.1093/brain/awm052 (visited on 05/25/2020).

[123] D. M. Clark and A. Wells, "A cognitive model of social phobia," *Social phobia: Diagnosis, assessment, and treatment*, vol. 41, no. 68, pp. 00 022–3, 1995.

[124] M. E. Coles and R. G. Heimberg, "Memory biases in the anxiety disorders: Current status," eng, *Clinical Psychology Review*, vol. 22, no. 4, pp. 587–627, May 2002, ISSN: 0272-7358. DOI: 10.1016/s0272-7358(01)00113-1.

[125] E. L. Daleiden, "Childhood Anxiety and Memory Functioning: A Comparison of Systemic and Processing Accounts," en, *Journal of Experimental Child Psychology*, vol. 68, no. 3, pp. 216–235, Mar. 1998, ISSN: 0022-0965. DOI: 10.1006/jecp.1997.2429. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022096597924292 (visited on 08/05/2020).

[126] A. Richards, "Anxiety and the resolution of ambiguity," in *Cognition, Emotion and Psychopathology: Theoretical, Empirical and Clinical Directions*, J. Yiend, Ed., Cambridge: Cambridge University Press, 2004, pp. 130–148, ISBN: 978-0-521-83391-2. DOI: 10.1017/CBO9780511521263.008. [Online]. Available: https://www.cambridge.org/core/books/cognition-emotion-and-psychopathology/anxiety-and-the-resolution-of-ambiguity/7C7E6560474389283CE98AB8C9393DE2 (visited on 08/05/2020).

[127] McClure, E.B., Adler, A., Monk, C.S., Cameron, J., Smith, S., Nelson, E.E., Leibenluft, E., Ernst, M. and Pine, D.S., "Fmri predictors of treatment outcome in pediatric anxiety disorders," *Psychopharmacology*, vol. 191, no. 1, pp. 97–105, 2007.

[128] H. A. Marusak, J. M. Carré, and M. E. Thomason, "The stimuli drive the response: An fmri study of youth processing adult or child emotional face stimuli," *NeuroImage*, vol. 83, pp. 679–689, 2013.

[129] M. Taddei, E. Bertoletti, A. Zanoni, and M. Battaglia, "Responses to facial expressions of emotions in social anxiety disorder: A review of fmri and erp studies from a developmental psychopathology perspective.," *Mind & Brain, The Journal of Psychiatry*, vol. 1, no. 2, 2010.

[130] K. Beesdo, S. Knappe, and D. S. Pine, "Anxiety and anxiety disorders in children and adolescents: Developmental issues and implications for DSM-V," en, *Psychiatric Clinics of North America*, vol. 32, no. 3, pp. 483–524, Sep. 2009, ISSN: 0193953X. DOI: 10.1016/j.psc.2009.06.002. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0193953X09000562 (visited on 08/05/2020).

[131] K. M. Thomas, "Amygdala response to fearful faces in anxious and depressed children," en, *Archives of General Psychiatry*, vol. 58, no. 11, pp. 1057–1063, Nov. 2001, ISSN: 0003990X. DOI: 10.1001/archpsyc.58.11.1057. [Online]. Available: http://archpsyc.ama-assn.org/cgi/doi/10.1001/archpsyc.58.11.1057 (visited on 03/28/2020).

[132] A. Frick, K. Howner, H. Fischer, M. Kristiansson, and T. Furmark, "Altered fusiform connectivity during processing of fearful faces in social anxiety disorder," en, *Translational Psychiatry*, vol. 3, no. 10, e312–e312, Oct. 2013, ISSN: 2158-3188. DOI: 10.1038/tp.2013.85. [Online]. Available: http://www.nature.com/articles/tp201385 (visited on 08/06/2020).

[133] Fonzo, G.A., Ramsawh, H.J., Flagan, T.M., Sullivan, S.G., Letamendi, A., Simmons, A.N., Paulus, M.P. and Stein, M.B., "Common and disorder-specific neural responses to emotional faces in generalized anxiety, social anxiety and panic disorders," *The British Journal of Psychiatry*, vol. 206, no. 3, pp. 206–215, Mar. 2015, ISSN: 0007-1250. DOI: 10.1192/bjp.bp.114.149880. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4345308/ (visited on 05/29/2020).

[134] Fu, C. H., Mourao-Miranda, J., Costafreda, S. G., Khanna, A., Marquand, A. F., Williams, S. C., Brammer, M. J., "Pattern classification of sad facial processing: Toward the development of neurobiological markers in depression," en, *Biological Psychiatry*, vol. 63, no. 7, pp. 656–662, Apr. 2008, ISSN: 00063223. DOI: 10.1016/j.biopsych.2007.08.020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0006322307008773 (visited on 03/25/2020).

[135] Liu, F., Guo, W., Fouche, J.P., Wang, Y., Wang, W., Ding, J., Zeng, L., Qiu, C., Gong, Q., Zhang, W. and Chen, H., "Multivariate classification of social anxiety disorder using whole brain functional connectivity," en, *Brain Structure and Function*, vol. 220, no. 1, pp. 101–115, Jan. 2015, ISSN: 1863-2653, 1863-2661. DOI: 10.1007/s00429-013-0641-4. [Online]. Available: http://link.springer.com/10.1007/s00429-013-0641-4 (visited on 03/28/2020).

[136] Yuen, K.S., Johnston, S.J., De Martino, F., Sorger, B., Formisano, E., Linden, D.E. and Goebel, R., "Pattern classification predicts individuals' responses to affective stimuli," en, *Translational Neuroscience*, vol. 3, no. 3, Jan. 2012, ISSN: 2081-6936, 2081-3856. DOI: 10.2478/s13380-012-0029-6. [Online]. Available: http://www.degruyter.com/view/

j/tnsci.2012.3.issue-3/s13380-012-0029-6/s13380-012-0029-6.xml (visited on 03/28/2020).

[137] J. V. Haxby, A. C. Connolly, and J. S. Guntupalli, "Decoding neural representational spaces using multivariate pattern analysis," en, *Annual Review of Neuroscience*, vol. 37, no. 1, pp. 435–456, Jul. 2014, ISSN: 0147-006X, 1545-4126. DOI: 10.1146/annurev-neuro-062012-170325. [Online]. Available: http://www.annualreviews.org/doi/10.1146/annurev-neuro-062012-170325 (visited on 04/02/2020).

[138] M. Khosla, K. Jamison, G. H. Ngo, A. Kuceyeski, and M. R. Sabuncu, "Machine learning in resting-state fmri analysis," *Magnetic resonance imaging*, vol. 64, pp. 101–121, 2019.

[139] Xu, J., Van Dam, N.T., Feng, C., Luo, Y., Ai, H., Gu, R. and Xu, P., "Anxious brain networks: A coordinate-based activation likelihood estimation meta-analysis of resting-state functional connectivity studies in anxiety," en, *Neuroscience & Biobehavioral Reviews*, vol. 96, pp. 21–30, Jan. 2019, ISSN: 01497634. DOI: 10.1016/j.neubiorev.2018.11.005. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0149763418304123 (visited on 08/11/2020).

[140] Kim, M.J., Loucks, R.A., Palmer, A.L., Brown, A.C., Solomon, K.M., Marchante, A.N. and Whalen, P.J., "The structural and functional connectivity of the amygdala: From normal emotion to pathological anxiety," *Behavioural brain research*, vol. 223, no. 2, pp. 403–410, Oct. 2011, ISSN: 0166-4328. DOI: 10.1016/j.bbr.2011.04.025. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3119771/ (visited on 08/11/2020).

[141] S. L. Rauch, L. M. Shin, and C. I. Wright, "Neuroimaging studies of amygdala function in anxiety disorders," en, *Annals of the New York Academy of Sciences*, vol. 985, no. 1, pp. 389–410, 2003, ISSN: 1749-6632. DOI: 10.1111/j.1749-6632.2003.tb07096.x. [Online]. Available: https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.2003.tb07096.x (visited on 08/11/2020).

[142] A. Z. Snyder and M. E. Raichle, "A brief history of the resting state: The Washington University perspective," *Neuroimage*, vol. 62, no. 2, pp. 902–910, Aug. 2012, ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2012.01.044. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3342417/ (visited on 08/11/2020).

[143] M. E. Raichle, "A Paradigm Shift in Functional Brain Imaging," *The Journal of Neuroscience*, vol. 29, no. 41, pp. 12 729–12 734, Oct. 2009, ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.4366-09.2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6665302/ (visited on 08/11/2020).

[144] D. S. Pine, A. E. Guyer, and E. Leibenluft, "Functional magnetic resonance imaging and pediatric anxiety," *Journal of the American Academy of Child and Adolescent Psychiatry*, vol. 47, no. 11, pp. 1217–1221, Nov. 2008, ISSN: 0890-8567. DOI: 10.1097/CHI.0b013e318185dad0. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2635373/ (visited on 04/23/2020).

[145] Carpenter, K.L., Angold, A., Chen, N.K., Copeland, W.E., Gaur, P., Pelphrey, K., Song, A.W. and Egger, H.L., "Preschool anxiety disorders predict different patterns of amygdala-prefrontal connectivity at school-age," en, *PLOS ONE*, vol. 10, no. 1, A. Kavushansky, Ed., e0116854, Jan. 2015, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0116854. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0116854 (visited on 05/28/2020).

[146] K. L. H. Carpenter, A. Angold, N.-K. Chen, *et al.*, *"preschool anxiety disorders"*, OpenNeuro, 2018. DOI: null.

[147] Egger, H.L., Erkanli, A., Keeler, G., Potts, E., Walter, B.K. and Angold, A., "Test-Retest reliability of the Preschool Age Psychiatric Assessment (PAPA)," eng, *Journal of the American Academy of Child and Adolescent Psychiatry*, vol. 45, no. 5, pp. 538–549, May 2006, ISSN: 0890-8567. DOI: 10.1097/01.chi.0000205705.71194.b8.

[148] Angold, A., Prendergast, M., Cox, A., Harrington, R., Simonoff, E. and Rutter, M., "The Child and Adolescent Psychiatric Assessment (CAPA)," eng, *Psychological Medicine*, vol. 25, no. 4, pp. 739–753, Jul. 1995, ISSN: 0033-2917. DOI: 10.1017/s003329170003498x.

[149] W. H. Organization, Ed., *International classification of functioning, disability and health: ICF*, en. Geneva: World Health Organization, 2001, OCLC: ocm48250012, ISBN: 978-92-4-154542-6.

[150] O. Esteban, C. J. Markiewicz, R. W. Blair, *et al.*, "fMRIPrep: A robust preprocessing pipeline for functional MRI," *Nature Methods*, vol. 16, no. 1, pp. 111–116, Jan. 2019, ISSN: 1548-7105. DOI: 10.1038/s41592-018-0235-4. [Online]. Available: https://www.nature.com/articles/s41592-018-0235-4 (visited on 07/20/2020).

[151] M. Yousefnezhad, A. Selvitella, D. Zhang, A. J. Greenshaw, and R. Greiner, "Shared space transfer learning for analyzing multi-site fmri data," *arXiv preprint arXiv:2010.15594*, 2020.

[152] M. Jenkinson, M. Pechaud, and S. Smith, "BET2 - MR-Based Estimation of Brain, Skull and Scalp Surfaces," en, p. 1,

[153] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," eng, *NeuroImage*, vol. 17, no. 2, pp. 825–841, Oct. 2002, ISSN: 1053-8119. DOI: 10.1016/s1053-8119(02)91132-8.

[154] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009, ISBN: 1441412697.

[155] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[156] P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.

[157] W. McKinney *et al.*, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, Austin, TX, vol. 445, 2010, pp. 51–56.

[158] C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, 2020. DOI: 10.1038/s41586-020-2649-2.

[159] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.

[160] H. Popal, Y. Wang, and I. R. Olson, "A guide to representational similarity analysis for social neuroscience," *Social cognitive and affective neuroscience*, vol. 14, no. 11, pp. 1243–1253, 2019.

[161] N. Kriegeskorte, M. Mur, and P. A. Bandettini, "Representational similarity analysis-connecting the branches of systems neuroscience," *Frontiers in systems neuroscience*, vol. 2, p. 4, 2008.

[162] M. B. Cai, N. W. Schuck, J. W. Pillow, and Y. Niv, "Representational structure or task structure? Bias in neural representational similarity analysis and a Bayesian method for reducing bias," en, *PLOS Computational Biology*, vol. 15, no. 5, S. Jbabdi, Ed., e1006299, May 2019, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006299. [Online]. Available: https://dx.plos.org/10.1371/journal.pcbi.1006299 (visited on 07/05/2020).

[163] M. Cai, N. W. Schuck, J. W. Pillow, and Y. Niv, "A bayesian method for reducing bias in neural representational similarity analysis," *Advances in Neural Information Processing Systems*, vol. 29, pp. 4951–4959, 2016.

[164] R. W. Cox, "Afni: Software for analysis and visualization of functional magnetic resonance neuroimages," *Computers and Biomedical research*, vol. 29, no. 3, pp. 162–173, 1996.

[165] Z. S. Saad, R. C. Reynolds, B. Argall, S. Japee, and R. W. Cox, "Suma: An interface for surface-based intra-and inter-subject analysis with afni," in *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)*, IEEE, 2004, pp. 1510–1513.

[166] Chen, P.H.C., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J. and Ramadge, P.J., "A Reduced-Dimension fMRI Shared Response Model," en, p. 9, 2015.

[167] Zeng, L.L., Shen, H., Liu, L., Wang, L., Li, B., Fang, P., Zhou, Z., Li, Y. and Hu, D., "Identifying major depression using whole-brain functional connectivity: A multivariate pattern analysis," en, *Brain*, vol. 135, no. 5, pp. 1498–1507, May 2012, ISSN: 1460-2156, 0006-8950. DOI: 10.1093/brain/aws059. [Online]. Available: https://academic.oup.com/brain/article-lookup/doi/10.1093/brain/aws059 (visited on 08/11/2020).

[168] C. Wong and J. Gallate, "The function of the anterior temporal lobe: A review of the empirical evidence," en, *Brain Research*, vol. 1449, pp. 94–116, Apr. 2012, ISSN: 00068993. DOI: 10.1016/j.brainres.2012.02.017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0006899312002946 (visited on 06/29/2020).

[169] W. K. Simmons and A. Martin, "The anterior temporal lobes and the functional architecture of semantic memory," en, *Journal of the International Neuropsychological Society*, vol. 15, no. 5, pp. 645–649, Sep. 2009, ISSN: 1355-6177, 1469-7661. DOI: 10.1017/S1355617709990348. [Online]. Available: https://www.cambridge.org/core/product/identifier/S1355617709990348/type/journal_article (visited on 06/29/2020).

[170] P. Gloor, A. Olivier, L. F. Quesney, F. Andermann, and S. Horowitz, "The role of the limbic system in experiential phenomena of temporal lobe epilepsy," en, *Annals of Neurology*, vol. 12, no. 2, pp. 129–144, 1982, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ana.410120203, ISSN: 1531-8249. DOI: 10.1002/ana.410120203. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.410120203 (visited on 05/25/2020).

[171] M.-M. Mesulam, *Principles of Behavioral and Cognitive Neurology*, en. Oxford University Press, Jan. 2000, Google-Books-ID: kezqJb69OlAC, ISBN: 978-0-19-803080-5.

[172] K. Nakamura and K. Kubota, "The primate temporal pole: Its putative role in object recognition and memory," en, *Behavioural Brain Research*, vol. 77, no. 1-2, pp. 53–77, May 1996, ISSN: 01664328. DOI: 10.1016/0166-4328(95)00227-8. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0166432895002278 (visited on 05/25/2020).

[173] C. G. Gross and J. Sergent, "Face recognition," en, *Current Opinion in Neurobiology*, vol. 2, no. 2, pp. 156–161, Apr. 1992, ISSN: 0959-4388. DOI: 10.1016/0959-4388(92)90004-5. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0959438892900045 (visited on 05/27/2020).

[174] D. Perrett and D. Potter, "Visual cells in the temporal cortex sensitive to face view and gaze direction," en, *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 223, no. 1232, pp. 293–317, Jan. 1985, ISSN: 0080-4649, 2053-9193. DOI: 10.1098/rspb.1985.0003. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rspb.1985.0003 (visited on 05/27/2020).

[175] H. Kondo, K. S. Saleem, and J. L. Price, "Differential connections of the temporal pole with the orbital and medial prefrontal networks in macaque monkeys," eng, *The Journal of Comparative Neurology*, vol. 465, no. 4, pp. 499–523, Oct. 2003, ISSN: 0021-9967. DOI: 10.1002/cne.10842.

[176] ——, "Differential connections of the perirhinal and parahippocampal cortex with the orbital and medial prefrontal networks in macaque monkeys," eng, *The Journal of Comparative Neurology*, vol. 493, no. 4, pp. 479–509, Dec. 2005, ISSN: 0021-9967. DOI: 10.1002/cne.20796.

[177] Dougherty, D.D., Shin, L.M., Alpert, N.M., Pitman, R.K., Orr, S.P., Lasko, M., Macklin, M.L., Fischman, A.J. and Rauch, S.L., "Anger in healthy men: A PET study using script-driven imagery," eng, *Biological Psychiatry*, vol. 46, no. 4, pp. 466–472, Aug. 1999, ISSN: 0006-3223. DOI: 10.1016/s0006-3223(99)00063-3.

[178] Phillips, M.L., Young, A.W., Scott, S., Calder, A.J., Andrew, C., Giampietro, V., Williams, S.C., Bullmore, E.T., Brammer, M. and Gray, J.A., "Neural responses to facial and vocal expressions of fear and disgust," eng, *Proceedings. Biological Sciences*, vol. 265, no. 1408, pp. 1809–1817, Oct. 1998, ISSN: 0962-8452. DOI: 10.1098/rspb.1998.0506.

[179] G. Glosser, A. E. Salvucci, and N. D. Chiaravalloti, "Naming and recognizing famous faces in temporal lobe epilepsy," en, *Neurology*, vol. 61, no. 1, pp. 81–86, Jul. 2003, Publisher: Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology Section: Articles, ISSN: 0028-3878, 1526-632X. DOI: 10.1212/01.WNL.0000073621.18013.E1. [Online]. Available: https://n.neurology.org/content/61/1/81 (visited on 05/27/2020).

[180] T. Tsukiura, M. Namiki, T. Fujii, and T. Iijima, "Time-dependent neural activations related to recognition of people's names in emotional and neutral face-name associative learning: An fMRI study," eng, *NeuroImage*, vol. 20, no. 2, pp. 784–794, Oct. 2003, ISSN: 1053-8119. DOI: 10.1016/S1053-8119(03)00378-1.

[181] Devlin, J.T., Russell, R.P., Davis, M.H., Price, C.J., Wilson, J., Moss, H.E., Matthews, P.M. and Tyler, L.K., "Susceptibility-induced loss of signal: Comparing PET and fMRI on a semantic task," eng, *NeuroImage*, vol. 11, no. 6 Pt 1, pp. 589–600, Jun. 2000, ISSN: 1053-8119. DOI: 10.1006/nimg.2000.0595.

[182] R. J. Davidson and B. S. McEwen, "Social influences on neuroplasticity: Stress and interventions to promote well-being," en, *Nature Neuroscience*, vol. 15, no. 5, pp. 689–695, May 2012, ISSN: 1097-6256, 1546-1726. DOI: 10.1038/nn.3093. [Online]. Available: http://www.nature.com/articles/nn.3093 (visited on 03/25/2020).

[183] L. A. Miller, K. H. Taber, G. O. Gabbard, and R. A. Hurley, "Neural Underpinnings of Fear and Its Modulation: Implications for Anxiety Disorders," en, *The Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 17, no. 1, pp. 1–6, Feb. 2005, ISSN: 0895-0172, 1545-7222. DOI: 10.1176/jnp.17.1.1. [Online]. Available: http://psychiatryonline.org/doi/abs/10.1176/jnp.17.1.1 (visited on 05/28/2020).

[184] R. J. Dolan and P. Vuilleumier, "Amygdala automaticity in emotional processing," eng, *Annals of the New York Academy of Sciences*, vol. 985, pp. 348–355, Apr. 2003, ISSN: 0077-8923. DOI: 10.1111/j.1749-6632.2003.tb07093.x.

[185] J. E. LeDoux, "Emotion circuits in the brain," eng, *Annual Review of Neuroscience*, vol. 23, pp. 155–184, 2000, ISSN: 0147-006X. DOI: 10.1146/annurev.neuro.23.1.155.

[186] A. Kling and H. D. Steklis, "A neural substrate for affiliative behavior in nonhuman primates," eng, *Brain, Behavior and Evolution*, vol. 13, no. 2-3, pp. 216–238, 1976, ISSN: 0006-8977. DOI: 10.1159/000123811.

[187] Fullana, M.A., Harrison, B.J., Soriano-Mas, C., Vervliet, B., Cardoner, N., Àvila-Parcet, A. and Radua, J., "Neural signatures of human fear conditioning: An updated and extended meta-analysis of fMRI studies," en, *Molecular Psychiatry*, vol. 21, no. 4, pp. 500–508, Apr. 2016, Number: 4 Publisher: Nature Publishing Group, ISSN: 1476-5578. DOI: 10.1038/mp.2015.88. [Online]. Available: https://www.nature.com/articles/mp201588 (visited on 06/17/2020).

[188] J. E. LeDoux, *Anxious: Using the brain to understand and treat fear and anxiety*. Penguin, 2015.

[189] L. Pessoa and R. Adolphs, "Emotion processing and the amygdala: From a'low road'to'many roads' of evaluating biological significance," *Nature reviews neuroscience*, vol. 11, no. 11, pp. 773–782, 2010.

[190] R. Adolphs, "Fear, faces, and the human amygdala," *Current opinion in neurobiology*, vol. 18, no. 2, pp. 166–172, 2008.

[191] P. J. Whalen, "The uncertainty of it all," *Trends in cognitive sciences*, vol. 11, no. 12, pp. 499–500, 2007.

[192] L. F. Barrett, "The theory of constructed emotion: An active inference account of interoception and categorization," *Social cognitive and affective neuroscience*, vol. 12, no. 1, pp. 1–23, 2017.

[193] L. Pessoa, "A network model of the emotional brain," *Trends in cognitive sciences*, vol. 21, no. 5, pp. 357–371, 2017.

[194] Veer, I.M., Beckmann, C., Van Tol, M.J., Ferrarini, L., Milles, J., Veltman, D., Aleman, A., Van Buchem, M.A., Van Der Wee, N.J. and Rombouts, S.A., "Whole brain resting-state analysis reveals decreased functional connectivity in major depression," *Frontiers in Systems Neuroscience*, vol. 4, Sep. 2010, ISSN: 1662-5137. DOI: 10.3389/fnsys.2010.00041. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2950744/ (visited on 08/11/2020).

[195] Eugene, F., Lévesque, J., Mensour, B., Leroux, J.M., Beaudoin, G., Bourgouin, P. and Beauregard, M., "The impact of individual differences on the neural circuitry underlying sadness," en, *NeuroImage*, vol. 19, no. 2, pp. 354–364, Jun. 2003, ISSN: 1053-8119. DOI: 10.1016/S1053-8119(03)00121-6. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811903001216 (visited on 06/29/2020).

[196] Lévesque, J., Eugene, F., Joanette, Y., Paquette, V., Mensour, B., Beaudoin, G., Leroux, J.M., Bourgouin, P. and Beauregard, M., "Neural circuitry underlying voluntary suppression of sadness," English, *Biological Psychiatry*, vol. 53, no. 6, pp. 502–510, Mar. 2003, Publisher: Elsevier, ISSN: 0006-3223, 1873-2402. DOI: 10.1016/S0006-3223(02)01817-6. [Online]. Available: https://www.biologicalpsychiatryjournal.com/article/S0006-3223(02)01817-6/abstract (visited on 06/29/2020).

[197] R. J. R. Blair, J. S. Morris, C. D. Frith, D. I. Perrett, and R. J. Dolan, "Dissociable neural responses to facial expressions of sadness and anger," en, *Brain*, vol. 122, no. 5, pp. 883–893, May 1999, ISSN: 1460-2156, 0006-8950. DOI: 10.1093/brain/122.5.883. [Online]. Available: https://academic.oup.com/brain/article-lookup/doi/10.1093/brain/122.5.883 (visited on 06/29/2020).

[198] Kimbrell, T.A., George, M.S., Parekh, P.I., Ketter, T.A., Podell, D.M., Danielson, A.L., Repella, J.D., Benson, B.E., Willis, M.W., Herscovitch, P. and Post, R.M., "Regional brain activity during transient self-induced anxiety and anger in healthy adults," en, *Biological Psychiatry*, vol. 46, no. 4, pp. 454–465, Aug. 1999, ISSN: 00063223. DOI: 10.1016/S0006-3223(99)00103-1. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0006322399001031 (visited on 06/27/2020).

[199] Blair, K., Shaywitz, J., Smith, B.W., Rhodes, R., Geraci, M., Jones, M., McCaffrey, D., Vythilingam, M., Finger, E., Mondillo, K. and Jacobs, M., "Response to emotional expressions in generalized social phobia and generalized anxiety disorder: Evidence for separate disorders," en, *American Journal of Psychiatry*, vol. 165, no. 9, pp. 1193–1202, Sep. 2008, ISSN: 0002-953X, 1535-7228. DOI: 10.1176/appi.ajp.2008.07071060.

[Online]. Available: http://psychiatryonline.org/doi/abs/10.1176/appi.ajp.2008.07071060 (visited on 05/29/2020).

[200] H.-J. Yoon, E. H. Seo, J.-J. Kim, and I. H. Choo, "Neural correlates of self-referential processing and their clinical implications in social anxiety disorder," *Clinical Psychopharmacology and Neuroscience*, vol. 17, no. 1, p. 12, 2019.

[201] Li, W., Cui, H., Zhu, Z., Kong, L., Guo, Q., Zhu, Y., Hu, Q., Zhang, L., Li, H., Li, Q. and Jiang, J., "Aberrant functional connectivity between the amygdala and the temporal pole in drug-free generalized anxiety disorder," en, *Frontiers in Human Neuroscience*, vol. 10, Nov. 2016, ISSN: 1662-5161. DOI: 10.3389/fnhum.2016.00549. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnhum.2016.00549/full (visited on 06/29/2020).

[202] K. McRae, J. J. Gross, J. Weber, *et al.*, "The development of emotion regulation: An fmri study of cognitive reappraisal in children, adolescents and young adults," *Social cognitive and affective neuroscience*, vol. 7, no. 1, pp. 11–22, 2012.

[203] K. P. Rankin, J. H. Kramer, P. Mychack, and B. L. Miller, "Double dissociation of social functioning in frontotemporal dementia," *Neurology*, vol. 60, no. 2, pp. 266–271, 2003.

[204] H. Klumpp and N. Amir, "Examination of vigilance and disengagement of threat in social anxiety with a probe detection task," *Anxiety, Stress & Coping*, vol. 22, no. 3, pp. 283–296, 2009.

[205] N. Amir, C. T. Taylor, J. A. Bomyea, and C. L. Badour, "Temporal allocation of attention toward threat in individuals with posttraumatic stress symptoms," *Journal of anxiety disorders*, vol. 23, no. 8, pp. 1080–1085, 2009.

[206] B. E. Gibb, C. A. Schofield, and M. E. Coles, "Reported history of childhood abuse and young adults' information-processing biases for facial displays of emotion," *Child maltreatment*, vol. 14, no. 2, pp. 148–156, 2009.

[207] A. Fernández-Jaén, S. López-Martın, J. Albert, *et al.*, "Cortical thinning of temporal pole and orbitofrontal cortex in medication-naive children and adolescents with adhd," *Psychiatry Research: Neuroimaging*, vol. 224, no. 1, pp. 8–13, 2014.

[208] N. Boddaert, N. Chabane, H. Gervais, *et al.*, "Superior temporal sulcus anatomical abnormalities in childhood autism: A voxel-based morphometry mri study," *Neuroimage*, vol. 23, no. 1, pp. 364–369, 2004.

[209] G. C. Patton, C. Coffey, H. Romaniuk, *et al.*, "The prognosis of common mental disorders in adolescents: A 14-year prospective cohort study," *The Lancet*, vol. 383, no. 9926, pp. 1404–1411, 2014.

[210] E. I. Martin, K. J. Ressler, E. Binder, and C. B. Nemeroff, "The Neurobiology of Anxiety Disorders: Brain Imaging, Genetics, and Psychoneuroendocrinology," en, *Psychiatric Clinics of North America*, vol. 32, no. 3, pp. 549–575, Sep. 2009, ISSN: 0193953X. DOI: 10.1016/j.psc.2009.05.004. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0193953X09000501 (visited on 03/25/2020).

[211] C. Benjet, E. Bromet, E. Karam, *et al.*, "The epidemiology of traumatic event exposure worldwide: Results from the world mental health survey consortium," *Psychological medicine*, vol. 46, no. 2, pp. 327–343, 2016.

[212] V. M. Bridgland, E. K. Moeck, D. M. Green, *et al.*, "Why the covid-19 pandemic is a traumatic stressor," *Plos one*, vol. 16, no. 1, e0240146, 2021.

[213] F. S. Alshehri, Y. Alatawi, B. S. Alghamdi, A. A. Alhifany, and A. Alharbi, "Prevalence of post-traumatic stress disorder during the covid-19 pandemic in saudi arabia," *Saudi Pharmaceutical Journal*, vol. 28, no. 12, pp. 1666–1673, 2020.

[214] S. Canada, "Survey on covid-19 and mental health,september to december 2020," *The Daily*, vol. 11, no. 001, 2021.

[215] P. Tucker and C. Czapla, "Post-covid stress disorder: Another emerging consequence of the global pandemic," *Psychiatric Times*, 2020.

[216] N. W. Chew, G. K. Lee, B. Y. Tan, *et al.*, "A multinational, multicentre study on the psychological outcomes and associated physical symptoms amongst healthcare workers during covid-19 outbreak," *Brain, behavior, and immunity*, vol. 88, pp. 559–565, 2020.

[217] Y.-X. Wang, H.-T. Guo, X.-W. Du, W. Song, C. Lu, and W.-N. Hao, "Factors associated with post-traumatic stress disorder of nurses exposed to corona virus disease 2019 in china," *Medicine*, vol. 99, no. 26, 2020.

[218] A. Kichloo, M. Albosta, K. Dettloff, *et al.*, "Telemedicine, the current covid-19 pandemic and the future: A narrative review and perspectives moving forward in the usa," *Family medicine and community health*, vol. 8, no. 3, 2020.

[219] N. Breslau, "Epidemiologic studies of trauma, posttraumatic stress disorder, and other psychiatric disorders," *The Canadian Journal of Psychiatry*, vol. 47, no. 10, pp. 923–929, 2002.

[220] M. Salehi, M. Amanat, M. Mohammadi, *et al.*, "The prevalence of post-traumatic stress disorder related symptoms in coronavirus outbreaks: A systematic-review and meta-analysis," *Journal of affective disorders*, 2021.

[221] S. Hong, H. Kim, and M. K. Park, "Impact of covid-19 on post-traumatic stress symptoms in the general population: An integrative review," *International Journal of Mental Health Nursing*, 2021.

[222] D. M. Hedderich and S. B. Eickhoff, "Machine learning for psychiatry: Getting doctors at the black box?" *Molecular psychiatry*, vol. 26, no. 1, pp. 23–25, 2021.

[223] Z. Zhou, T.-C. Wu, B. Wang, H. Wang, X. M. Tu, and C. Feng, "Machine learning methods in psychiatry: A brief introduction," *General psychiatry*, vol. 33, no. 1, 2020.

[224] E. C. Meltzer, T. Averbuch, J. H. Samet, *et al.*, "Discrepancy in diagnosis and treatment of post-traumatic stress disorder (ptsd): Treatment for the wrong reason," *The journal of behavioral health services & research*, vol. 39, no. 2, pp. 190–201, 2012.

[225] L. F. Ramos-Lima, V. Waikamp, T. Antonelli-Salgado, I. C. Passos, and L. H. M. Freitas, "The use of machine learning techniques in trauma-related disorders: A systematic review," *Journal of psychiatric research*, vol. 121, pp. 159–172, 2020.

[226] R. Rogers, K. W. Sewell, and A. M. Goldstein, "Explanatory models of malingering: A prototypical analysis.," *Law and human behavior*, vol. 18, no. 5, p. 543, 1994.

[227] M. Matto, D. E. McNiel, and R. L. Binder, "A systematic approach to the detection of false ptsd.," *The journal of the American Academy of Psychiatry and the Law*, vol. 47, no. 3, pp. 325–334, 2019.

[228] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *CoRR*, vol. abs/1708.05148, 2017. arXiv: `1708.05148`. [Online]. Available: `http://arxiv.org/abs/1708.05148`.

[229] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investigative Otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[230] J. M. Smyth, "Written emotional expression: Effect sizes, outcome types, and moderating variables.," *Journal of consulting and clinical psychology*, vol. 66, no. 1, p. 174, 1998.

[231] L. A. Gottschalk and G. C. Gleser, *The measurement of psychological states through the content analysis of verbal behavior*. Univ of California Press, 1979.

[232] C. L. Franklin and K. E. Thompson, "Response style and posttraumatic stress disorder (ptsd): A review," *Journal of Trauma & Dissociation*, vol. 6, no. 3, pp. 105–123, 2005.

[233] F. Ringeval, B. Schuller, M. Valstar, *et al.*, "Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 3–12.

[234] J. Gratch, R. Artstein, G. M. Lucas, *et al.*, "The distress analysis interview corpus of human and computer interviews.," in *LREC*, 2014, pp. 3123–3128.

[235] S. Scherer, G. M. Lucas, J. Gratch, A. S. Rizzo, and L.-P. Morency, "Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 59–73, 2015.

[236] D. DeVault, K. Georgila, R. Artstein, *et al.*, "Verbal indicators of psychological distress in interactive dialogue with a virtual human," in *Proceedings of the SIGDIAL 2013 Conference*, 2013, pp. 193–202.

[237] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency, "Automatic nonverbal behavior indicators of depression and ptsd: Exploring gender differences," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, IEEE, 2013, pp. 147–152.

[238] D. Banerjee, K. Islam, K. Xue, *et al.*, "A deep transfer learning approach for improved post-traumatic stress disorder diagnosis," *Knowledge and Information Systems*, vol. 60, no. 3, pp. 1693–1724, 2019.

[239] Q. He, B. P. Veldkamp, and T. de Vries, "Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach," *Psychiatry research*, vol. 198, no. 3, pp. 441–447, 2012.

[240] M. Oakes, R. Gaaizauskas, H. Fowkes, A. Jonsson, V. Wan, and M. Beaulieu, "A method based on the chi-square test for document classification," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 440–441.

[241] Q. He, B. P. Veldkamp, C. A. Glas, and T. de Vries, "Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining," *Assessment*, vol. 24, no. 2, pp. 157–172, 2017.

[242] V. Leiva and A. Freire, "Towards suicide prevention: Early detection of depression on social media," in *International Conference on Internet Science*, Springer, 2017, pp. 428–436.

[243] M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in *Proceedings of the 5th annual ACM web science conference*, 2013, pp. 47–56.

[244] J. Brubaker, F. Kivran-Swaine, L. Taber, and G. Hayes, "Grief-stricken in a crowd: The language of bereavement and distress in social media," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 6, 2012.

[245] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, 2014.

[246] L. Zhang, J. Driscol, X. Chen, and R. Hosseini Ghomi, "Evaluating acoustic and linguistic features of detecting depression sub-challenge dataset," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 47–53.

[247] D. DeVault, R. Artstein, G. Benn, *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 2014, pp. 1061–1068.

[248] J. A. Luno, M. Louwerse, and J. G. Beck, "Tell us your story: Investigating the linguistic features of trauma narrative," in *Proceedings of the annual meeting of the cognitive science society*, vol. 35, 2013.

[249] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, IEEE, vol. 1, 2006, pp. I–I.

[250] F. Morbini, D. DeVault, K. Sagae, J. Gerten, A. Nazarian, and D. Traum, "Flores: A forward looking, reward seeking, dialogue manager," in *Natural interaction with robots, knowbots and smartphones*, Springer, 2014, pp. 313–325.

[251] H. Brugman, A. Russel, and X. Nijmegen, "Annotating multi-media/multi-modal resources with elan.," in *LREC*, Citeseer, 2004.

[252] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "Flair: An easy-to-use framework for state-of-the-art nlp," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, 2019, pp. 54–59.

[253] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[254] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011, pp. 142–150.

[255] S. Loria, "Textblob documentation," *Release 0.15*, vol. 2, 2018.

[256] N. Vejkanchana and P. Kuacharoen, "Continuous variable binning algorithm to maximize information value using genetic algorithm," in *International Conference on Applied Informatics*, Springer, 2019, pp. 158–172.

[257] T. J. Cleophas and A. H. Zwinderman, "Optimal binning," in *Machine Learning in Medicine*, Springer, 2013, pp. 37–46.

[258] J. L. Lustgarten, V. Gopalakrishnan, H. Grover, and S. Visweswaran, "Improving classification performance with discretization on biomedical datasets," in *AMIA annual symposium proceedings*, American Medical Informatics Association, vol. 2008, 2008, p. 445.

[259] J. Sawalha, M. Yousefnezhad, A. M. Selvitella, B. Cao, A. J. Greenshaw, and R. Greiner, "Predicting pediatric anxiety from the temporal pole using neural responses to emotional faces," *Scientific Reports*, vol. 11, no. 1, pp. 1–17, 2021.

[260] E. C. Polley and M. J. Van Der Laan, "Super learner in prediction," 2010.

[261] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *The Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.

[262] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017. [Online]. Available: `http://jmlr.org/papers/v18/16-365.html`.

[263] A. G. Reece, A. J. Reagan, K. L. Lix, P. S. Dodds, C. M. Danforth, and E. J. Langer, "Forecasting the onset and course of mental illness with twitter data," *Scientific reports*, vol. 7, no. 1, pp. 1–11, 2017.

[264] C. R. Marmar, A. D. Brown, M. Qian, *et al.*, "Speech-based markers for posttraumatic stress disorder in us veterans," *Depression and anxiety*, vol. 36, no. 7, pp. 607–616, 2019.

[265] B. Sun, Y. Zhang, J. He, *et al.*, "A random forest regression method with selected-text feature for depression assessment," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 61–68.

[266] S. Suthaharan, "Support vector machine," in *Machine learning models and algorithms for big data classification*, Springer, 2016, pp. 207–235.

[267] R. Chiong, G. S. Budhi, and S. Dhakal, "Combining sentiment lexicons and content-based features for depression detection," *IEEE Intell. Syst*, vol. 36, no. 6, 2021.

[268] A. Kumar, A. Sharma, and A. Arora, "Anxious depression prediction in real-time social data," in *International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019, Uttaranchal University, Dehradun, India*, 2019.

[269] S. Saifullah, Y. Fauziah, and A. S. Aribowo, "Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data," *arXiv preprint arXiv:2101.06353*, 2021.

[270] J. H. Friedman, "Stochastic gradient boosting," *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[271] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," *Institute for Signal and information Processing*, vol. 18, no. 1998, pp. 1–8, 1998.

[272] C. for Substance Abuse Treatment *et al.*, "Understanding the impact of trauma," in *Trauma-informed care in behavioral health services*, Substance Abuse and Mental Health Services Administration (US), 2014.

[273] B. T. Litz, B. T. Litz, and M. J. Gray, "Emotional numbing in posttraumatic stress disorder: Current and future research directions," *Australian & New Zealand Journal of Psychiatry*, vol. 36, no. 2, pp. 198–204, 2002.

[274] A. Dell'Omo, *Trauma and recovery, by judith herman (1992)*, 2016.

[275] L. Goldfine, *Narrating Hurricane Katrina: Identifying linguistic patterns in survivors' trauma accounts*. City University of New York, 2010.

[276] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual review of psychology*, vol. 54, no. 1, pp. 547–577, 2003.

[277] R. S. Campbell and J. W. Pennebaker, "The secret life of pronouns: Flexibility in writing style and physical health," *Psychological science*, vol. 14, no. 1, pp. 60–65, 2003.

[278] J. Jaeger, K. M. Lindblom, K. Parker-Guilbert, and L. A. Zoellner, "Trauma narratives: It's what you say, not how you say it.," *Psychological Trauma: Theory, Research, Practice, and Policy*, vol. 6, no. 5, p. 473, 2014.

[279] M. A. Cohn, M. R. Mehl, and J. W. Pennebaker, "Linguistic markers of psychological change surrounding september 11, 2001," *Psychological science*, vol. 15, no. 10, pp. 687–693, 2004.

[280] Y. Son, S. A. Clouston, R. Kotov, *et al.*, "World trade center responders in their own words: Predicting ptsd symptom trajectories with ai-based language analyses of interviews," *Psychological Medicine*, pp. 1–9, 2011.

[281] J. Alvarez-Conrad, L. A. Zoellner, and E. B. Foa, "Linguistic predictors of trauma pathology and physical health," *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, vol. 15, no. 7, S159–S170, 2001.

[282] B. M. Wardecker, R. S. Edelstein, J. A. Quas, I. M. Cordón, and G. S. Goodman, "Emotion language in trauma narratives is associated with better psychological adjustment among survivors of childhood sexual abuse," *Journal of language and social psychology*, vol. 36, no. 6, pp. 628–653, 2017.

[283] E. B. Foa and B. O. Rothbaum, *Treating the trauma of rape: Cognitive-behavioral therapy for PTSD*. Guilford Press, 2001.

[284] B. Kleim, A. B. Horn, R. Kraehenmann, M. R. Mehl, and A. Ehlers, "Early linguistic markers of trauma-specific processing predict post-trauma adjustment," *Frontiers in psychiatry*, vol. 9, p. 645, 2018.

[285] B. A. Sharpless and J. P. Barber, "A clinician's guide to ptsd treatments for returning veterans.," *Professional Psychology: Research and Practice*, vol. 42, no. 1, p. 8, 2011.

[286] E. B. Foa, T. M. Keane, M. J. Friedman, and J. A. Cohen, *Effective treatments for PTSD: practice guidelines from the International Society for Traumatic Stress Studies*. Guilford Press, 2010.

[287] S. Papini, P. Yoon, M. Rubin, T. Lopez-Castro, and D. A. Hien, "Linguistic characteristics in a non-trauma-related narrative task are associated with ptsd diagnosis and symptom severity.," *Psychological Trauma: Theory, Research, Practice, and Policy*, vol. 7, no. 3, p. 295, 2015.

[288] M. J. Tanana, C. S. Soma, P. B. Kuo, *et al.*, "How do you feel? using natural language processing to automatically rate emotion in psychotherapy," *Behavior Research Methods*, pp. 1–14, 2021.

[289] J. Joormann and C. H. Stanton, "Examining emotion regulation in depression: A review and future directions," *Behaviour research and therapy*, vol. 86, pp. 35–49, 2016.

[290] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and combining sentiment analysis methods," in *Proceedings of the first ACM conference on Online social networks*, 2013, pp. 27–38.

[291] P. J. Lang and M. M. Bradley, "Emotion and the motivational brain," *Biological psychology*, vol. 84, no. 3, pp. 437–450, 2010.

[292] R. Bar-On, "The bar-on emotional quotient inventory (eq-i): Rationale, description and summary of psychometric properties.," 2004.

[293] F. Wright, "Negotiating the therapeutic alliance: A relational treatment guide," *The Journal of Psychotherapy Practice and Research*, vol. 10, no. 2, p. 138, 2001.