Towards general mental health biomarkers: machine learning analysis of multi-disorder EEG data

AKSHAY TALEKAR

Thesis Submitted to the University of Ottawa
in partial fulfillment of the Requirements for the Master's of Science, Mathematics and Statistics, Specialization in Biostatistics

,

Department of Mathematics and Statistics
Faculty of Science
University Of Ottawa

Master's of Science (2023)
Department Of Mathematics and Statistics
University Of Ottawa
Ottawa, Ontario, Canada

TITLE: Towards general mental health biomarkers: machine learning analysis of multi-disorder EEG data

AUTHOR:
Akshay Talekar,
B.Sc. (Major in Biochemistry, Minor in Statistics)

SUPERVISOR:
Dr. Maia Fraser
Associate Professor, Department of Mathematics and Statistics,
University Of Ottawa, ON, Canada

THESIS EXAMINERS:
Dr. Kelly Burkett,
Associate Professor, Department of Mathematics and Statistics,
University Of Ottawa, ON, Canada

Dr. Shirley Mills
Associate Professor, School of Mathematics and Statistics,
Carleton University, ON, Canada

Abstract

Several studies have made use of EEG features to detect specific mental health illnesses such as epilepsy or schizophrenia, as supplementary diagnosis to the usual symptom-based diagnoses. At the same time general mental health diagnostic tools (biomarker or symptom-based) to identify individuals who are manifesting early signs of mental health disorders are not commonly available. This thesis seeks to explore the potential use of EEG features as a biomarker-based tool for general mental health diagnosis [8].

Specifically, the predictive ability using machine learning of a general biomarker derived from EEG readings elicited from an oddball auditory experiment to predict someone's mental health status (mentally ill or healthy) is investigated in this study. Given that mindfulness exercises are regularly provided as treatment for a wide range of mental illnesses [14], the features of interest seek to quantify it as a measure of mental health. The 2 feature sets developed and tested in this study were collected from a traumatic brain injury (TBI) and healthy controls dataset [7]. Further testing of these feature sets was done on the Bipolar and Schizophrenia Network on Intermediate Phenotypes (BSNIP) dataset containing multiple mental illnesses and healthy controls [39] to test the features for generalizability. Feature Set 1 consisted of the average and variance of P300 and N200 ERP component peak amplitudes and latencies across the centro-parietal and fronto-central EEG channels respectively. Feature Set 2 contains the average and variance of P300 and N200 ERP component mean amplitudes across the centro-parietal and fronto-central EEG channels respectively.

The predictive ability of these 2 feature sets was tested. Logistic regression, support vector machines, decision trees, random forests, KNN classification algorithms were used, and random forest and KNN were used in combination with oversampling to predict the mental health status of the subjects (whether they were cases or healthy controls). The model performance was tested using accuracy, precision, sensitivity, specificity, f1 score, confusion matrices, and AUC of the ROC.

The results of this thesis show promise on the use of EEG features as biomarkers to diagnose mental illnesses or to get a better understanding of mental wellness. The use of this technology opens doors for more accurate, biomarker-based diagnosis of mental health conditions, lowering the cost of mental health care, and making mental health care accessible for more people [23].

Preface

This research was done for my thesis project as a part of my Master's degree in Science, Mathematics and Statistics, Specialization in Biostatistics at the University of Ottawa. Prior to completing my master's degree, I completed a Bachelor of Science, with a Major in Biochemistry and a Minor in Statistics from the University of Ottawa.

The inclination for this research work originally came through my own struggles with mental health during the COVID-19 pandemic. I often found myself questioning whether or not I should seek help. Looking for clarity for myself, I talked to my friends and family and many of them reported similar confusion with their mental health during and even before the pandemic. I then realized that this perplexity is common as many mental illnesses have overlapping symptoms and this can be hard for anyone to navigate due to lack of training and knowledge. This led me to the conclusion that there is a lack of resources for people to investigate and assess their own mental health conveniently.

The target group of the dissertation is researchers and professionals working in the fields of machine learning, psychology, and neuroscience. This thesis aims to inspire more research on the viability of using EEG features as biomarkers to test for disruptions in mental well-being in order to make diagnoses and thus early treatment more accessible.

This thesis would not have been possible without the thorough work done by researchers before me. This thesis was also made possible due to the contributions of the thesis supervisor Dr. Maia Fraser, and advice on datasets she solicited from Dr. Georg Northoff.

Acknowledgements

I would like to acknowledge my supervisor Dr. Maia Fraser, her guidance in the conceptualization and throughout the implementation of my thesis project.

Contents

List of Figures

List of Tables

Declaration of Academic Achievement

I, Akshay Talekar, declare that this thesis titled, **Towards general mental health biomarkers: machine learning analysis of multi-disorder EEG data**, and works presented in it are my own.

Introduction

It is estimated that in any given year, 1 in 5 Canadians experience a mental illness [35]. The burden of mental health illnesses is felt in all aspects of life. Almost 12 Canadians per day die by suicide [29]. It has also been estimated that worldwide, the disease burden for mental illness accounts for 32.4% of years lived with disability and 13% of disability-adjusted life-years [42]. Financially, the cost to the Canadian economy due to mental illness is estimated to be around 50 billion dollars [26]. The burden of a mental illness is immense but can be lowered with early detection. Early detection has many benefits such as preventing mental illness altogether, making treatments more effective, and giving time to patients to prepare and acquire the resources they need to make recovery easier [12], [25].

Mental illnesses are mainly diagnosed through a detailed interview using the DSM V (fifth edition of The Diagnostic and Statistical Manual of Mental Disorders). The DSM V provides a system to classify mental illnesses and allows mental health professionals to diagnose their patients. The DSM V does so by providing criteria and thresholds for mental illnesses [32]. Diagnoses made using the DSM V are based on clinical symptoms and these symptoms are required to meet certain thresholds to fit into the criteria sets for more than 70 disorders that the DSM V diagnoses [30]. While this method works adequately for diagnosing mental illnesses for people with symptoms that are clear, present, and meet the thresholds set by the DSM V, it does have drawbacks.

The DSM V has been widely criticized for not incorporating biomarkers into diagnosis; biomarkers would allow for the empirical testing of mental illnesses and reduce the number of false negatives [43]. The DSM V also does not help to indicate whether someone has a mental illness if they are presenting atypical symptoms, show symptoms that are not intense enough to meet the thresholds in the criteria sets of the DSM V, or are in the early stages of mental illness and have yet to develop symptoms [10], [41]. Just because someone is not diagnosed by

the DSM V criteria sets does not mean they are not suffering, and more information on personal mental health can be of immense help to people.

Given that mental health illnesses can present themselves in innumerable ways, a method of assessment for general mental health that is more subtle than the current DSM V criteria can be helpful to those suffering from symptoms that are not typical or intense enough to meet the thresholds set by the criteria sets of the DSM V [28].This thesis seeks to investigate a general biomarker from EEG readings that can be used for determining a subject's mental health illness status (case or control) for several illnesses. Such methods assess how the brain reacts to stimuli and based on that evidence, will see how well they discern whether someone is mentally ill. .

The development of a biomarker started by investigating mindfulness. Mindfulness is beneficial to mental health and is used as a treatment for many mental illnesses such as anxiety, depression, and OCD (obsessive-compulsive disorder) [14]. Given that raising a patient's mindfulness is a large part of therapy, a reasonable assumption can be made that a person's ability to be mindful may be indicative of their mental health [9]. This thesis builds on this hypothesis.

Defining mindfulness is still an emerging area of research, but the main components of mindfulness include a person's attentional readiness. Features describing attentional readiness were calculated through the EEG readings used in this thesis [4]. The EEG readings for subjects in this study were collected using the auditory oddball experiments where a subject listens to non target tones interspersed with an oddball target tone. Once the EEG readings were collected, the many trials during an EEG reading were epoched, and the readings were averaged to get a measure called an ERP (Event Related Potential) [38]. An ERP has many features or components depending on the stimulus or activity that invoked the ERP. In an auditory oddball experiment, the most common features that are evoked are N200, and P300 components [34] .

An N200 component is defined as a negative spike in voltage in an ERP at the 200-350ms after the stimulus and is usually measured at the fronto-central channels and the ERP is associated with perceptual novelty. A P300 component is defined as a positive spike in voltage in an ERP at the 250-400ms after the stimulus and is usually measured at the centro-parietal channels and the ERP is associated with stimulus significance and novelty. Features such as amplitude, mean amplitude, and feature latency were extracted from this data. [38].

Supervised learning methods, specifically logistic regression, support vector machines (SVM), decision tree classifier, random forest classifier, and K-nearest neighbors (KNN), were used to assess the performance of two feature sets in predicting mental illness status (healthy control vs case). Supervised learning is a category of machine learning used in situations where one looks to predict an output variable based on a set of input variables and pre-labeled data is available [13]. In this case it was known which subject was mentally ill. Each example in supervised learning contained an input and an output. As examples from the training set (a portion, usually 70%,

of the whole dataset) are analyzed by the algorithm, its weights are adjusted for an optimal fit. Once the model is fitted or trained, the performance was tested using the test set by using the model to predict the outputs for the test set and comparing the predicted values to the actual values. [13]

The performance of the trained models were then tested using a variety of methods. The accuracy of the prediction is measured, and other measures such as true positive rate, false positive rate, true negative rate, and false negative rate are calculated as well as the F1 score [19]. A receiver operating characteristic curve is also used to compare the model to a random classifier. A training method called cross-validation is also implemented. Cross-validation iteratively trains the model by re sampling different portions of the data as opposed to using just one training and testing set [31], [37].

Getting treatment early in case of mental illness can help the patients by preventing more serious illness later and help ease the burden on the mental health system [12], [25]. Helping patients understand their mental health better can also allow them access to treatment or other assistance (such as housing, disability assistance etc.). This method of assessment using EEG data is particularly useful for disorders such as PTSD where interviewing a patient can cause harm to the patient as they are forced to relive the trauma [15]. Instead, the patient's EEG data can be assessed by having them perform a completely unrelated task. Due to the relatively low cost and wide availability of EEG technology, such an assessment tool can be used effectively to screen patients and be of use in underdeveloped or remote communities where mental health resources are scarce [23].

Literature Review

2.1 Introduction:

EEG has been widely used to measure neural activity for academic and commercial purposes. Research has shown that EEG biomarkers can be used to predict specific mental illnesses in people. However, in order to thoroughly research applications of EEG biomarkers to predict general mental well-being, it is important to first investigate prior work done in the areas of EEG biomarkers, and the role of mindfulness in mental well-being. This literature review aims to explore the current body of research on EEG biomarkers for predicting mental illnesses, the role of mindfulness in mental well being, and to identify gaps in the literature.

2.2 Methodology:

A comprehensive search of relevant electronic databases, search engines and academic repositories (e.g. PubMed, PsycINFO, Google Scholar, NIMH) was conducted to identify relevant articles published between 2010 and 2020. Some examples of search terms used were "EEG Biomarkers", "Mental illness", "predicting mental illness with EEG", "mindfulness and mental health", "quantifying mindfulness". The selected studies were limited to research conducted on humans, published in English, and with a focus on predicting mental illness and the role of mindfulness in mental health.

2.3 Results:

A total of 15 studies were included in this literature review. The studies cover many mental illnesses such as depression, anxiety, schizophrenia, bipolar disorder. EEG biomarkers used in

these papers were of different categories including spectral power, connectivity measures, event-related potentials (ERPs), and phase-amplitude coupling (PAC).

The results of the studies suggest that EEG biomarkers can be used to predict specific mental illnesses with moderate to high accuracy.

For example, one study used Relative Wavelet Energy (RWE) and artificial neural network (ANN) classifier to predict Major depressive disorders from healthy controls with an accuracy of 98.11% [2].

Similar examples can also be found with schizophrenia, one study used 17 features which included neural oscillations (beta, delta waves etc.), correlation dimension (CD), entropy features (sample entropy, approximate entropy etc.), these 17 features were then used as inputs for logistic regression (LR), SVMs, KNN, decision trees (DT), and random forests (RF) giving classification accuracies 89% for SVM, 87% for RF , 86% for LR , 86% for KNN and 68% for DT [33].

One study aims to predict ADHD, the study consisted of 181 patients with ADHD and 147 healthy controls. This study made use of spectral power and ERP amplitude and latency measures to classify ADHD with sensitivity ranging from 75% to 83% and specificity values of 71% to 77% [27].

The efforts made to predict specific illnesses in this literature review are summarized below:

TABLE 2.1: Studies predicting specific mental illnesses using EEG Biomarkers.

Author	Illness	Features	Methods	Results
Puthankattil et al, 2012	Major Depressive Disorder	RWE, signal entropy	ANN	acc = 98.11%
Muller et al., 2019	ADHD	spectral power, ERP amplitude and latency	Logistic Regression	Sensitivity = 75% - 83% , Specificity = 71% - 77%.
Ahmadlou et al., 2012	Major Depressive Disorder	KFD,HFD	PNN Classifier	Accuracy = 91.3%
Slimen et al., 2020	Epilepsy	Spike Rate detection	Statistical Analysis	Accuracy = 92%
Poil et al., 2013	Alzheimer at the MCI stage	28 EEG Biomarkers	logistic regression	sensitivity of 88% and specificity of 82%

Behnam et al., 2016	epilepsy	Interpolated histogram feature	Distribution model, Bayesian classifier, Hunting search , MLP	Accuracy = 86.56%
Rodrigues et al., 2016	Alzheimers	NMax, NMin, Zcr, Mdif, RP	Statistical Analysis	ROC AUC = 0.934, sens= 86.19%, spec = 99.35%, acc = 94.88%
Giannakakis et al., 2015	Stress/Anxiety state	Absolute power, relative power, coherence, symmetry	Statistical Analysis	N/A
Prabhakar et. al., 2020	Schizophrenia	9 biomarkers	SVM-RBFkernel	classification accuracy of 92.17%
Haveman et al., 2019	TBI severity	12 Biomarkers	Random Forest classifier	AUC = 0.84 (sensitivity 88%, specificity 73%)
Hosseinifard et al., 2013	Major Depressive Disorder	Higuchi fractal, correlation metrics, DFA, LLE	KNN, discriminant analysis, logistic regression	Accuracy = 90%
Handojoseno et al., 2013	Freezing of Gait in Parkinson's Disease patients	spatial, spectral, and temporal features of the EEG signals utilizing wavelet coefficients	Multilayer Perceptron Neural Network and k-Nearest Neighbor classifier	87% sensitivity and 73% accuracy

Betrouni et al., 2018	Parkinson's Disease	Relative power low electrode density	SVM, KNN	acc = 84% and 88%
Acharya et al., 2015	Major Depressive disorder	FD, LLE, Sample entropy, DFA, Hurst's exponent, HOS, RQA	SVM,KNN, DT, NC, PNN	Accuracy = 98%
Vanneste et al., 2018	Parkinson's Disease	Power Spectra of Bands and regions	SVM	Accuracy =94.34%

These studies suggest that EEG biomarkers can be used to predict specific mental illnesses with moderate to high accuracy. However, they do not bring forth a single set of features that is capable of capturing multiple mental illnesses at once. These studies focus on biomarkers that are specific to certain illnesses, not mental wellbeing as a whole. These studies also do not show an underlying common traits between multiple mental illnesses.

The studies also highlight some limitations of using EEG biomarkers for predicting general mental illness status. For example, the studies use different EEG features and analytical methods, making it difficult to compare results across studies. Additionally, the studies often had small sample sizes and were conducted on specific populations, which may limit generalizability of the findings.

2.4 Conclusion:

Overall, the literature suggests that different EEG biomarkers have been developed for specific illnesses, but more research is needed to identify reliable and valid EEG biomarkers for general mental health and to test their generalizability to larger and more diverse populations. Additionally, more standardized analytical methods are needed to facilitate comparisons across EEG biomarker studies or to make generalizations about commonality between mental illnesses. EEG biomarkers for general mental health have the potential to provide important insights into the mechanisms underlying mental illness, to inform the development of treatment for people with mental illness, and provide accessible mental health resources.

Methodology

All computations and analysis in this thesis was done using Python 3.7 using Jupyter Notebook. Libraries used are Pandas, Numpy, Imblearn, Sklearn, and MNE.

3.1 Data Source:

In this study EEG technology was chosen as the data type as it provides an instant reading, is widely available, portable, safe, and is not cost-prohibitive compared to other brain imaging technologies such as an MRI. EEG recordings also allow for the extraction of specific features that can be used to measure mindfulness through measuring attentional readiness. These features were tested in this study as a biomarker on their ability to predict mental illness status.

In this study, two datasets were used. The first is more restricted in the number of subjects and types of mental illnesses. This dataset was downloaded from OpenNeuro and was created by Dr. James F Cavanagh and Dr. Davin Quinn [7]. The data was collected from 2016 to 2018 in the Center for Brain Recovery and Repair at the University of New Mexico Health Sciences Center. This dataset contained EEG readings for an auditory oddball experiment for subjects who were either healthy controls or had a traumatic brain injury (TBI). Each subject had 3 sessions of recording, only session 1 was used as many subjects had sessions 2 and 3 missing due to subject attrition over time. Session 1 was recorded 3 to 14 days post-injury [7]. This dataset was used to develop the EEG features which would then be used to predict the subject's case status. In the TBI dataset all data was collection at one centre. Information on subject recruitment strategies was not available. All data was collected at one centre. This dataset has been used by the creator of the dataset in studies assessing difference in executive function after TBI [5] and a joint analysis of frontal theta synchrony and white matter [6]. This suggests that the dataset was intended to be used for a comparative study between controls and subjects with TBI.

The second dataset used was more expansive in terms of the number of subjects and types of mental illnesses. This dataset was downloaded from the NIMH (National Institute of Mental Health) data archive. This dataset is a part of a study called BNSIP (Bipolar and Schizophrenia Network on Intermediate Phenotypes). The goal of BSNIP dataset is due to the overlap in symptoms in many psychotic and mood disorders, to develop classifications of mental illnesses based on biological measurements, which may be more effective at identifying causes and treatments [39]. BSNIP is a joint effort between the Bio-imaging Research Center (BIRC) at the University of Georgia, Commonwealth Research Center, Harvard Medical School, University of Chicago Medical Center's Department of Psychiatry and Behavioral Neuroscience, University of Texas Southwestern Medical Center, The Olin Neuropsychiatry Research Center at the Institute of Living [39]. The BSNIP dataset contains EEG readings, fMRI data, Positive and Negative Symptom Scale (PANSS), Montgomery Asberg Depression Rating Scale, and Young Mania Rating Scale for healthy controls, people with mood disorders such as depression, bipolar disorder, and psychotic disorders such as schizophrenia [39]. The subject data in the BSNIP dataset was collected at multiple centres, however collection site information by subject was not available. Information on subject recruitment strategies was also not available.

The TBI and BSNIP datasets were chosen because the datasets needed to contain high-quality EEG data as well as clinical data about subjects such as their mental illness status (i.e. diagnoses or rating scales). Additional demographic data such as age, sex, race, socio-economic status were also of interest. Ideally, datasets would contain 500 subjects or more, this was achieved with the BSNIP dataset. For the purposes of generalizability, it was determined that at least one of the datasets would be needed containing controls as well as subjects diagnosed with multiple mental illnesses, the BSNIP dataset also met this requirement. This would ensure that the general biomarker being assessed is not just discerning between someone having only one mental illness (ex. Predicting if someone has depression or not) or not but discerning general mental health.

3.2 Data Cleaning:

3.2.1 TBI Dataset:

The data from the TBI dataset was imported into Python. Using metadata files, the response variable was determined to be "HadHeadInjury" which denoted using 0 and 1 if a subject has had a traumatic brain injury (TBI). This file had information on 93 unique subjects. There was no missing data in these columns. The TBI dataset contained EEG readings for 96 subjects, information on why this number of subjects differed from the number of unique subjects found in the metadata file was not available. After performing an inner join on the file paths for

subjects and meta data bases off of the study's unique identifier, 69 subjects remained who had both metadata and EEG readings. Of the 69 subjects, 44 had a traumatic brain injury and 25 were healthy controls. Of these 69 subjects 46 (67%) were randomly selected to the training set and the remaining 23 (33%) were the testing set. The relatively few healthy controls and the low final number of subjects with more EEG reading and metadata could be explained subject attrition. Heavy subject attrition was noted by the dataset creator on the dataset OpenNeuro page, however further information on subject attrition is not available (ex. original number of recruited subjects and their case status). Information on subject recruitment strategies was not available.

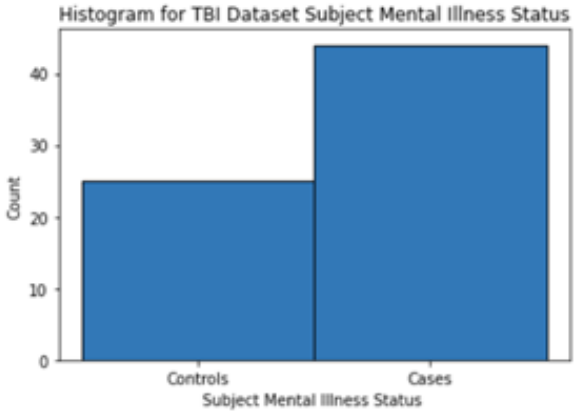


FIGURE 3.1: Histogram of TBI dataset subject mental illness status.

3.2.2 BSNIP Dataset:

The response variable was determined by inspecting the metadata .txt files downloaded with the B-SNIP data package. These files contained all data on the study subjects which was not EEG data. The response variable chosen for the B-SNIP dataset was labeled "Phenotype/diagnosis for the subject" and denoted whether the subject was a case (diagnosed with a mental illness) or a control. This response variable was chosen because it gave an objective and binary classification for a subject's mental illness status. This response variable is also similar to the response variable found in the TBI dataset.

Once the response variable was determined, the percentage of missing data per column in the metadata file (the file where the response variable was found) was calculated and it was seen that most columns either had all of their data, ie. 0% of rows with a value of NaN, or had close to none of it, ie. $> 90\%$ rows with a value of NaN. The decision was then made to drop columns that have 90% of the rows missing data as dropping the rows which have missing data for that column would result in massive loss of data.

The metadata file contained a unique identifier for each subject which was used to determine the number of duplicates. Out of the 4403 subjects in this file, 1988 were unique subjects. These duplicate subjects had the same value for the response variable, but with different values for dataset ID and collection ID (both variables which are not considered in the analysis performed in this thesis), therefore these duplicates were removed from the dataset by keeping the first instance of the duplicate. This large amount of duplicates can be explained by multiple visits caused due to the multiple reading measures acquired in this study.

The unique subjects were then merged with their EEG recording file path by using their unique identifier. An inner joined was performed to only receive subject who has both, metadata and EEG recordings. There were no duplicates in the EEG recording files. As a result of this inner join, 613 unique subjects' EEG readings and a response variable denoting their mental health status remain.

During the feature extraction step, a number of subjects were also removed for not containing the EEG electrodes needed for feature extraction, problems with file reading, and being outlier which would not justify an accurate EEG reading. The final number of subjects for which EEG features and a response variable were available after this data cleaning is 515, with 188 being healthy controls, and 327 being cases. Information on recruitment strategies was not available. Of these 515 subjects, 345 (67%) were randomly selected to the training set and the remaining 170 (33%) were the testing set.

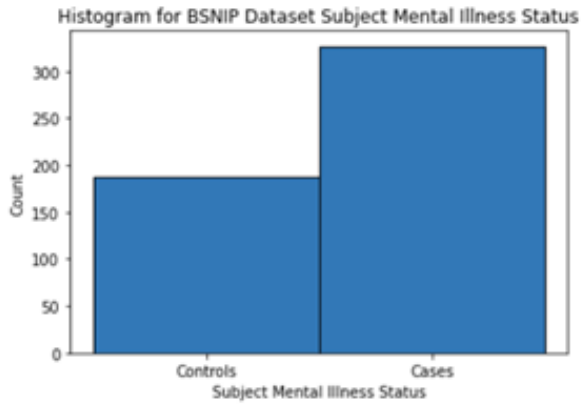
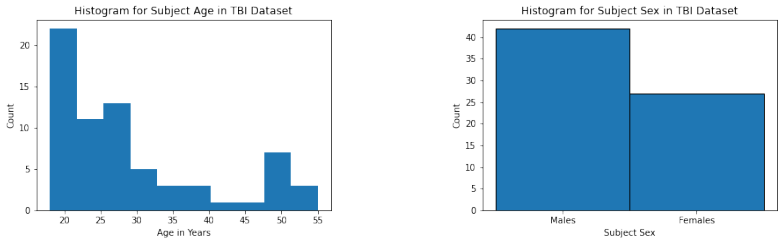


FIGURE 3.2: Histogram of BSNIP dataset subject mental illness status.

3.3 Exploratory Analysis:

3.3.1 TBI:

Exploratory data analysis for the TBI dataset was done by plotting the distributions of cases to controls which can be seen in Figure 3.1. The distributions of the metadata such as age, and sex were also plotted. It is seen in Figure 3.3 that there are more males than females, and that there is a bimodal distribution for age which slightly skews to the left (younger) side and has another peak at around 50 years.



(A) Histogram for Age in Years (B) Histogram for Sex at Birth

FIGURE 3.3: Distribution of Metadata in TBI Dataset

In Table 3.1, the T-test done for age shows us that age of subjects is not significantly different between cases and controls in the BSNIP dataset. The χ^2 test done for sex at birth is also significant at $\alpha = 0.05$. This shows that sex at birth distribution is also not significantly different cases and controls.

TABLE 3.1: Statistical test for differences in Metadata between Cases and Controls in TBI Dataset

	Mean Age (SD)	t	Sex at Birth - Count	χ^2
Cases	27.84 (9.87)	-1.284	F: 14 M: 30	2.73
Controls	31.44 (11.87)		F:13 M:12	

Tables 3.2 shows the descriptive statistics for Feature Set 1. This table shows the mean,with standard deviation, minimum and maximum values for each independent variable for both feature sets. In Feature Set 1, relatively high levels of standard deviations compared to the means are seen for the amplitude independent variables when compared to the latency variables in

Feature Set 1. There is also a larger difference between the minimum and maximum values in the amplitude independent variables compared to the latency variables in Feature Set 1.

TABLE 3.2: Descriptive statistics for TBI Feature Set 1

	P300 Amp Avg	P300 Amp Var	P300 Lat Avg	P300 Lat Var
Mean	0.000023	1.527254e-07	0.318285	0.001352
Std Dev	0.000172	1.267861e-06	0.027360	0.001058
Min	-0.000004	1.466788e-13	0.264000	0.000093
Max	0.001432	1.053173e-05	0.384333	0.004361

(A) P300

	N200 Amp Avg	N200 Amp Var	N200 Lat Avg	N200 Lat Var
Mean	-7.847770e-06	4.217165e-11	0.300995	1.145887e-03
Std Dev	3.796075e-06	2.832436e-10	0.036891	1.399742e-03
Min	-1.717101e-05	2.369554e-13	0.203333	3.081488e-33
Max	7.178659e-07	2.358973e-09	0.350000	4.570222e-03

(B) N200

In table 3.3 which shows descriptive statistics for Feature Set 2. In Feature Set 2, it can be seen that the P300 mean amplitude variance has very high values for mean, minimum and maximum. It is also see that the N200 Mean amplitude variance's sample standard deviance is very high. Additionally the P300 Mean amplitude average's sample standard deviation is also high.

TABLE 3.3: Descriptive statistics for TBI Feature Set 2

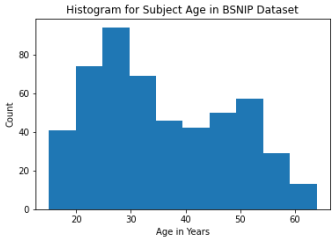
	P300 MA Avg	P300 MA Var	N200 MA Avg	N200 MA Var
Mean	19.542311	1.532610e+05	-3.002944	42.444468
Std Dev	170.373986	1.272455e+06	2.985794	319.926176
Min	-6.708142	1.279492e-01	-10.825427	0.068641
Max	1413.898132	1.056988e+07	9.388162	2661.245715

MA referring to Mean Amplitude.

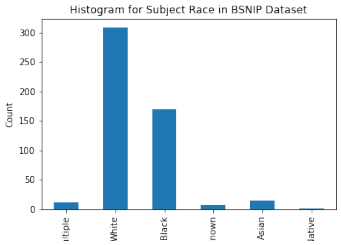
3.3.2 BSNIP:

Exploratory data analysis for the BSNIP dataset was done by first getting the distributions of cases to controls which can be seen in Figure 3.2, after that, the distributions of the metadata

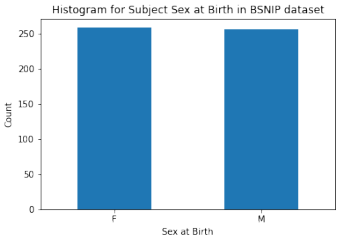
such as age, sex, and race were also visualized. It is seen in Figure 3.4 that there is a balance in sex, most subjects are white or black, and that there is a bimodal distribution for age which slightly skews to the left (younger).



(A) Histogram for Age in Years



(B) Histogram for Race



(C) Histogram for Sex at Birth

FIGURE 3.4: Distribution of Metadata in BSNIP Dataset

In Table 3.4, the T-test done for age shows us that age of subjects is significanly different between cases and controls in the BSNIP dataset. The χ^2 test done for sex at birth is also significant at $\alpha = 0.05$. This shows that sex at birth distribution is also significantly different cases and controls. Race was not found to be significant.

TABLE 3.4: Statistical test for differences in Metadata between Cases and Controls in BSNIP Dataset

	Mean Age (SD)	t	Sex at Birth - Count	χ^2	Race	χ^2
Cases	34.05 (12.29)	3.602*	F: 149 M: 178	8.01*	Black: 114 White: 192 Other: 21	1.59
Controls	38.13 (12.42)		F:110 M:78		Black: 56 White: 117 Other: 15	

Tables 3.5 shows the descriptive statistics for Feature Set 1. This table shows the mean,with

standard deviation, minimum and maximum values for each independent variable for both Feature Sets. In Feature Set 1, high levels of standard deviations are seen for the amplitude independent variables when compared to the latency variables in Feature Set 1. There is also a larger difference between the minimum and maximum values in the amplitude independent variables compared to the latency variables in Feature Set 1.

TABLE 3.5: Descriptive statistics for BSNIP Feature Set 1

	P300 Amp Avg	P300 Amp Var	P300 Lat Avg	P300 Lat Var
Mean	0.000119	9.282724e-07	0.349393	0.000822
Std Dev	0.001020	1.565223e-05	0.038856	0.001234
Min	-0.000077	2.373843e-14	0.250000	0.000000
Max	0.020132	3.440068e-04	0.400000	0.005137

(A) P300

	N200 Amp Avg	N200 Amp Var	N200 Lat Avg	N200 Lat Var
Mean	-0.000067	8.292068e-07	0.237914	0.000746
Std Dev	0.000613	1.465661e-05	0.031485	0.001089
Min	-0.012584	5.664196e-14	0.200000	0.000000
Max	0.000021	3.269596e-04	0.348000	0.004706

(B) N200

In table 3.6 which shows descriptive statistics for Feature set 2. In Feature Set 2, it can be seen that the standard deviation is proportionate to the mean for all independent variables. The difference between minimum and maximum values is also much smaller Feature Set 2.

TABLE 3.6: Descriptive statistics for BSNIP Feature Set 2

	P300 MA Avg	P300 MA Var	N200 MA Avg	N200 MA Var
Mean	0.6086	0.619	0.1111	0.9286
Std Dev	0.6522	0.6364	0.1111	1
Min	0.6087	0.6875	0.4444	0.6429
Max	0.6522	0.6923	0.5555	0.7857

MA referring to Mean Amplitude.

3.4 Feature Extraction:

The features to be extracted were determined through findings found in previous research.

In the field of psychiatry and mental health therapy, a common treatment for many mental illnesses is mindfulness exercises (ex. Guided meditation, progressive muscle relaxation, body scan, etc.) [18]. An increase in mindfulness has been shown to help treat patients suffering from mental illness [18]. Due to well-known therapeutic relationship, it begs the question if a person's ability to be mindful can predict someone's mental illness status.

Defining mindfulness is an ongoing task but many studies agree that mindfulness can be defined through an amalgamation of attentional readiness and mental coherence [4]. Attentional readiness can be measured by using the P300 and N200 components in an event related potential (ERP) which can be elicited using an auditory oddball task [1]. P300 and N200 components of an ERP can be used to measure attentional readiness as the P300 component is associated with stimulus significance and novelty and the N200 component is associated with perceptual novelty. [3]

P300 and N200 components' amplitudes and latencies are of importance, as changes in these are found in subjects with some mental illness [38]. A P300 component is defined as a positive spike in voltage in an ERP at the 250-400ms after the stimulus usually measured at the centro-parietal channels and an N200 component is defined as a negative spike in voltage in an ERP at the 200-350ms after the stimulus usually measured at the fronto-central channels [38].

Both, the TBI and BSNIP dataset contains data for 63 EEG channels, each corresponding to a specific brain area follow the 10-10 international EEG channel placement system. In this thesis, the six centro-pareital and the six fronto-central channels are of interest.

Graph 3.5 and 3.6 are examples of TBI subject 16's P300 and N200 ERP components respectively. Each plot has the respective 6 channels for P300 and N200 ERP component and the shaded region shows the area where P300 and N200 occur. The star shows the channel with the peak amplitude.

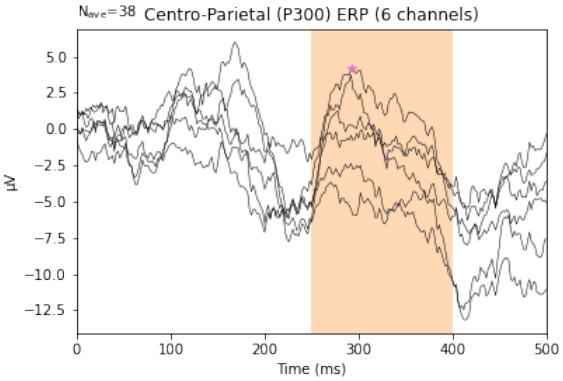


FIGURE 3.5: TBI Subject 16's ERP for P300 Component's 6 Centro-Pareital Channels.

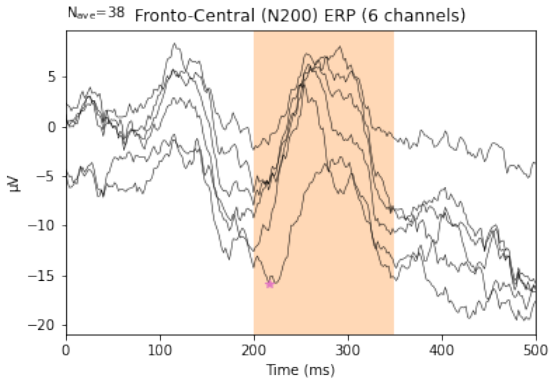


FIGURE 3.6: TBI Subject 16's ERP for N200 Component's 6 Fronto-Central Channels.

The following feature sets were to be extracted in order to be tested for their ability to predict mental illness:

- The first set of features:
  - Average of P300 Amplitude (averaged across 6 centro-parietal channels)
  - Variance of P300 Amplitude (variance calculated across 6 centro-parietal channels)
  - Average of P300 Latency (averaged across 6 centro-parietal channels)
  - Variance of P300 Latency (variance calculated across 6 centro-parietal channels)
  - Average of N200 Amplitude (averaged across 6 fronto-central channels)
  - Variance of N200 Amplitude (variance calculated across 6 fronto-central channels)
  - Average of N200 Latency (averaged across 6 fronto-central channels)
  - Variance of N200 Latency (variance calculated across 6 fronto-central channels)
- The second set of features:
  - Average of P300 Mean amplitude (averaged across 6 centro-parietal channels)
  - Variance of P300 Mean amplitude (variance calculated across 6 centro-parietal channels)
  - Average of N200 Mean amplitude (averaged across 6 fronto-central channels)
  - Variance of N200 Mean amplitude (variance calculated across 6 fronto-central channels)

The first set of features is being calculated at each of the 6 centro-parietal and 6 fronto-central. The readings are averages and variances are calculated across the respective channels to reduce the dimensionality of the features. The dimensionality is reduced given that the number of subjects in the TBI dataset is low, and a model input with high dimensionality would result in an inaccurate model.

The second set of features calculates the mean amplitude across the 6 channels of each respective component. The mean amplitude for the P300 component is calculated 250-400ms after the stimulus and the mean amplitude for the N200 component is calculated 200-350ms after the stimulus. Then, the average and variance for the mean amplitudes are calculated across the 6 channels. In Feature Set 2, latency is missing due to the mean amplitude being calculated over the entirety of the length of the component. The benefit of using such a feature is that calculating the mean amplitude acts as a low pass filter to increase the signal-to-noise ratio on the ERP reading.

Calculations necessary to extract the P300 and N200 features from the raw EEG readings are done slightly differently between the TBI dataset and the BSNIP dataset.

3.4.1 TBI Dataset:

In the TBI dataset, in order to extract the features listed above from the EEG readings elicited by an auditory oddball experiment, an ERP must first be calculated for each EEG reading [38]. In order to do this, only the target tone trials from the raw EEG reading are taken using the event code "S200" as this is the code associated with the target tone for the auditory oddball experiment. For the TBI dataset, there are 38 target tones per subject, each of these 38 trials containing a target tone is called an epoch.

Graph 3.7 lays out the first 5 epochs for Subject 001 of the TBI dataset, there are a total of 38 such epochs for the TBI dataset. The x-axis is the epoch, while the y-axis is the EEG channel, not all channels from the dataset are represented. Each subject has 46 channels:

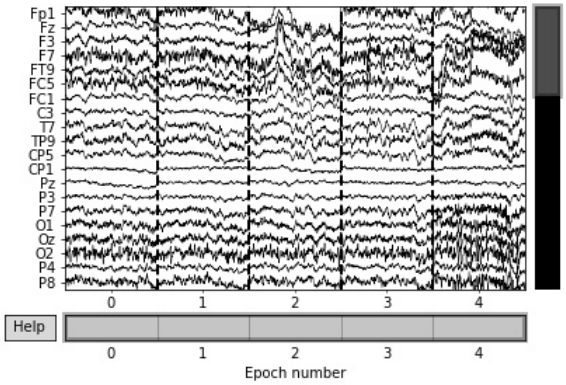


FIGURE 3.7: First 5 EEG reading epochs for various channels for TBI subject 001.

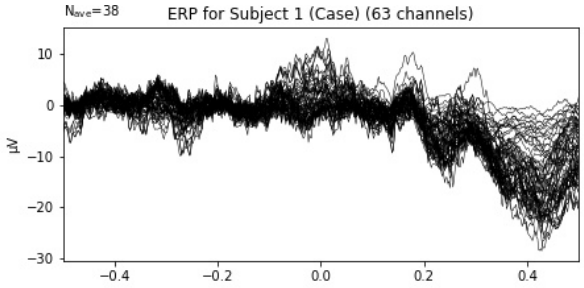


FIGURE 3.8: ERP for subject 1 containing all EEG channels

Once the raw EEG is epoched, the epochs are averaged to give the ERP. Figure 3.8 is an example of an ERP calculated for Subject 001 (who has a traumatic brain injury) in the TBI dataset for all 63 channels, the ERP was calculated by averaging over 38 epochs and contains data from 0.5s before the target stimulus event to 0.5 after the target stimulus event.

3.4.2 BSNIP Dataset:

In the BSNIP dataset, similarly, the raw EEG file is read and the data is loaded. The BSNIP dataset EEG readings do not contain EKG and VEOG channels. The raw EEG reading is then epoched using the event code "2" as this is the code associated with the target tone for the auditory oddball experiment. Once the raw EEG is epoched, the epochs are averaged to give the ERP.

3.5 Frameworks/Predictions:

The algorithms in this thesis make use of training dataset and a testing dataset for both the TBI and BSNIP datasets. Tables 3.7 and 3.8 give the distributions of the cases and controls for

the training sets, testing sets and the datasets as wholes.

TABLE 3.7: Number of cases and controls in the training and testing sets for the TBI dataset

	Training Set (67%)	Testing Set (33%)	Total
Control	13	12	25
Case	33	11	44
Total	46	23	69

TABLE 3.8: Number of cases and controls in the training and testing sets for the BSNIP dataset

	Training Set (67%)	Testing Set(33%)	Total
Control	136	52	188
Case	209	118	327
Total	345	170	515

In order to test the predictive ability of the features that were extracted from the EEG readings from the auditory oddball experiments, the features will be used as inputs into machine learning algorithms in order to predict the desired outcome for each of the datasets. Since the output variable in both datasets were binary, the same algorithms were used in both cases.

Each algorithm was trained on both feature sets in an effort to test the features' ability to predict the outcome variable and compare the performance of the features to each other. Logistic regression, support vector machine/classifier (SVC), random forest classifier, decision tree classifier, K-nearest neighbors (KNN) classifier were used. After applying these algorithms, oversampling is applied to train the best performing algorithms to improve performance.

3.5.1 Logistic Regression

Firstly, logistic regression was used to predict the values of the binary response variable in both the TBI and BSNIP datasets. Logistic regression is a parametric statistical model sometimes also known as a logit model, which makes use of a model made up of a linear combination of coefficients and variables to predict the log odds of the binary outcome.

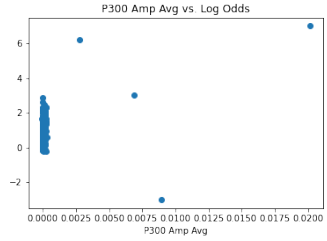
$$Logit(\pi) = \frac{1}{(1 + e^{-\pi})}$$

$$Ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 * X_1 + \ldots + \beta_k * X_k$$

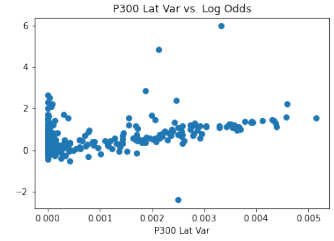
The beta parameters used in a logistic regression model can be estimated using maximum likelihood estimation. There are a few key assumptions that need to be met in order to employ the logistic regression algorithm to predict the response variables. The data needs to have an appropriate outcome type (binary), sufficiently large sample size, independent variable linearity with log odds, no multicollinearity between independent variables, no influential outliers, independence of observations. [36].

In order to perform modelling using logistic regression, first, the model needs to be fit on the training data, and then the model's parametric assumptions need to be evaluated to determine whether the model can be used.

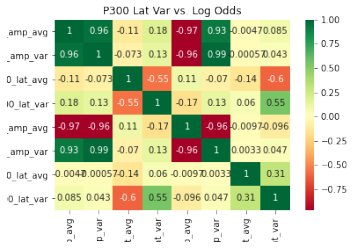
The model assumptions being checked in the following figures are for BSNIP dataset Feature Set 1.



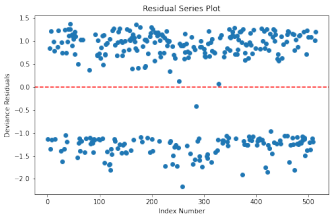
(A) P300 Amplitude Average vs Log Odds



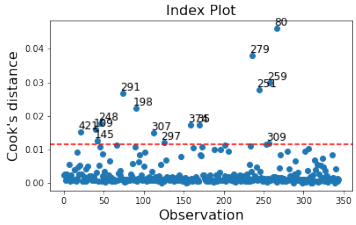
(B) P300 Latency Variance vs Log Odds



(C) Correlation Matrix for Feature Set 1



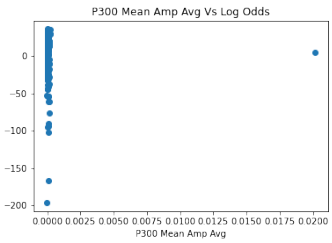
(D) Residuals Vs Index



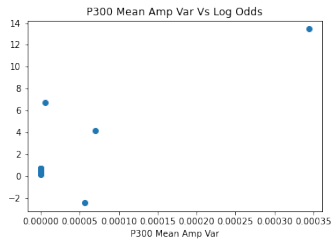
(E) Cook's Distance vs Index

FIGURE 3.9: Model Assumptions for Logistic Regression for BSNIP Feature Set 1.

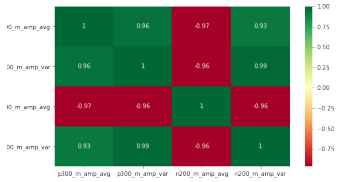
The same model assumptions were evaluated for BSNIP Feature Set 2.



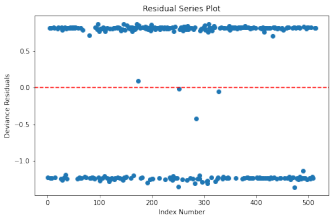
(A) P300 Mean Amplitude Average vs Log Odds



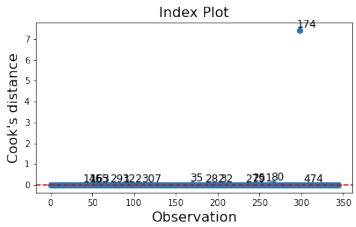
(B) P300 Mean Amplitude Variance vs Log Odds



(C) Correlation Matrix for Feature Set 2



(D) Residuals Vs Index



(E) Cook's Distance vs Index

FIGURE 3.10: Model Assumptions for Logistic Regression for BSNIP Feature Set 2.

The assumption of appropriate outcome variable and sufficient data are met, but it is seen from Figure 3.8 and 3.9 that the assumptions of linearity of independent variables with log odds is not met, the assumption of no or low multicollinearity is also not met, there are many influential outliers in Feature Set 1, while there is only one in Feature Set 2, independence of observations is seen in both feature sets.

Given that many of these assumptions to use the logistic regression algorithm are violated, the results from this algorithm may not be reliable. These graphs also show that data is very non linear, especially for Feature Set 1. This result was used to make the decision to not apply oversampling on the parametric methods (logistic regression and SVM), as oversampling would not do anything to affect the model assumptions and the results would still be unreliable.

3.5.2 Support Vector Machines:

The second algorithm used for classification was support vector machine (SVM). Support vector machines were selected as an algorithm because they are commonly used to perform classification of linear and nonlinear data. However, SVMs can be used for both classification and regression. SVM classify by mapping the input variables onto a high-dimensional space to maximize the width of the gap between the two or more categories, the algorithm then finds a hyperplane in that high dimensional space to classify different categories, for the purposes of classification many different hyperplanes can be chosen, however the objective is to maximize the distance between the two or more classes. These hyperplanes are used as decision boundaries to classify observations. The dimensions of the hyperplanes depend on the dimensions of the input data (ex. If the input has 2 dimensions, the hyperplane will have 1). Support vectors are data that are close to the hyperplane and influence the properties of the hyperplane, such as orientation. [11]

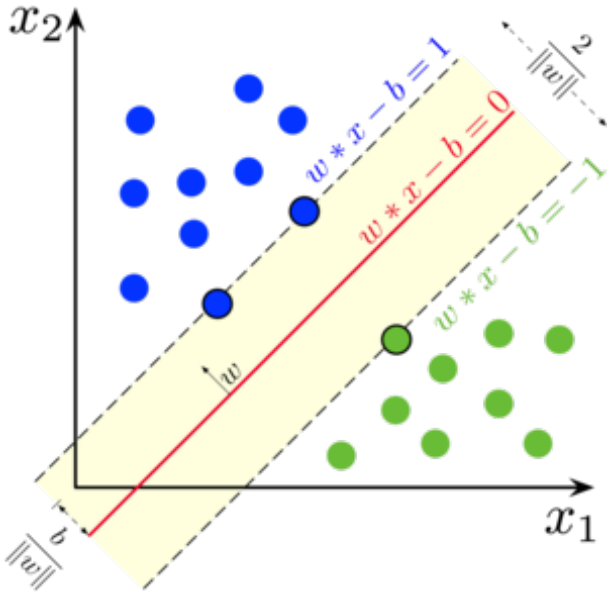


FIGURE 3.11: Depiction of maximum margin hyperplane and margins for SVM trained with observations of 2 classes. [44]

To apply the SVM algorithm on the 2 datasets and 2 feature sets, parameter tuning was necessary. the GridSearchCV function from the SKlearn library. This function tries all the combinations of defined parameter values and evaluates the model for each combination using the Cross-Validation method using a defined metric. The metric used to optimize model performance was the AUC of the ROC curve.

The parameters values tested by GridSearchCV:

- C: 0.01,0.1,1,10
- kernel: "linear","poly","rbf","sigmoid"
- degree : 1,3,5,7
- gamma' : 0.01,1

The parameters determined by that optimized ROC curve AUC are:

TABLE 3.9: Algorithm Parameters used for Support Vector Machines by Dataset and Feature Set.

Dataset	Feature Set	C	Kernel	Degree	Gamma
TBI	1	10	'rbf'	1	1
	2	0.1	'rbf'	1	1
BSNIP	1	10	'rbf'	1	1
	2	0.1	'rbf'	1	0.01

3.5.3 Decision Tree Classifiers:

Decision trees make use of a flowchart-like structure that branches off to predict different results as a consequence of certain decisions or input values. Decision tree algorithms are useful due to their high interpretability and as a result, are used in many business domains. A decision tree contains three types of nodes; decision nodes, chance nodes, and end nodes [22]. Some benefits of using decision trees are that they are easily interpretable, can be used with incomplete data, can be used to determine best and worst case scenarios, and reasons for outcome can be traced back to decisions. Decision tree algorithms are also non-parametric. Some drawbacks of using a decision tree algorithm are that they have high variance, meaning that they are sensitive to irregularities in training data which can cause them to be overtrained on the training data, causing low performance when deployed to a test set, decision trees are also relatively inaccurate when used not in combination with another training method such as ensemble learning or boosting. [22] . In order to address oversampling in the decision tree algorithm pre-pruning was performed in order to limit the tree depth. This pre-pruning was performed using GridSearchCV function in the SKlearn library in Python. Specifically, Pre-pruning was chosen as is it a more efficient method of pruning when compared to post-pruning which needs a decision tree to first be built to full depth before pruning as opposed to pre-pruning which limits the depth of the tree using model parameters.

The parameter values tested by GridSearchCV:

- criterion: “gini”, “entropy”, “log_loss”
- Max_depth: 3,5,10,15,20,None
- Min_samples_leaf : 1,2,5
- Min_samples_split : 2,5,7,10

The parameters determined that optimized ROC curve AUC are:

TABLE 3.10: Algorithm Parameters used for Decision Trees by Dataset and Feature Set.

Dataset	Feature Set	Criterion	Max_depth	Min_samples_leaf	Min_samples_split
TBI	1	'gini'	5	1	2
	2	'gini'	3	2	2
BSNIP	1	'gini'	3	1	2
	2	'gini'	3	5	2

3.5.4 Random Forest Classifiers:

The next classification algorithm used to differentiate cases from controls was a random forest classifier. Random forest classification makes use of ensemble learning. During the training of the model, random forest classifier builds many decision trees. Inputs are then classified into the category returned by most trees. Random forests generally outperform decision trees. Random forests averages multiple decision trees trained on different parts of the training set to reduce the variance [17]. While Random forest algorithms reduce overfitting compared to singular decision trees, overfitting can still happen especially case when trees get trained very deep. This is also known as having high variance. This problem can be solved by performing parameter tuning. Another benefit of random forest classifiers is that they are non-parametric.

In order to train on a set of data, random forest classifiers use a method called bagging. In bagging, the training algorithm randomly samples with replacement the training set and fits decision trees on these samples B times. After training, predictions for the test set can be made by taking the majority vote of all the trained decision trees. This training method makes the model less sensitive to irregularities or outliers in the data thus reducing the variance of the model. The random sampling with replacement makes sure that the decision trees are not highly correlated and that the same decision tree does not appear multiple times [17], [16]. The number of trees is proportional to the number of rows of the training set, the more rows there are, the more trees are trained. Similar to SVM and decision trees algorithms before, parameter

tuning to optimize ROC curve AUC for random forest classifier is also done using GridSearchCV from the SKlearn library in Python.

The parameter values tested by GridSearchCV:

- Max_depth: 2,3,5,10,20
- Max_features: 2,4,6,8
- Max_leaf_nodes: 2,5,7,10
- Min_samples_leaf: 5,10,20,50,100,200
- N_estimators: 10,25,30,50,100,200

The parameters determined that optimized ROC curve AUC are:

TABLE 3.11: Algorithm Parameters used for Random Forest Classifier by Dataset and Feature Set.

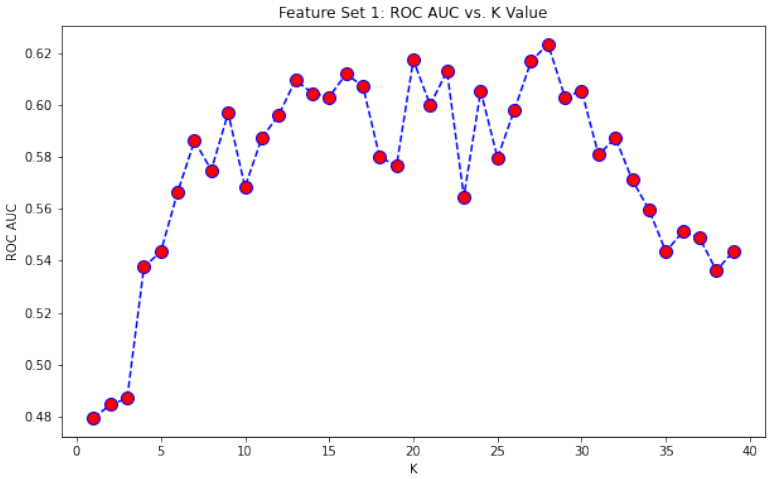
Dataset	Feature Set	Max depth	Max features	Max leaf nodes	Min samples leaf	N estimators
TBI	1	3	4	5	5	10
	2	2	4	2	5	10
BSNIP	1	10	4	10	5	25
	2	2	2	2	10	50

3.5.5 K-Nearest Neighbors (KNN) Classifier

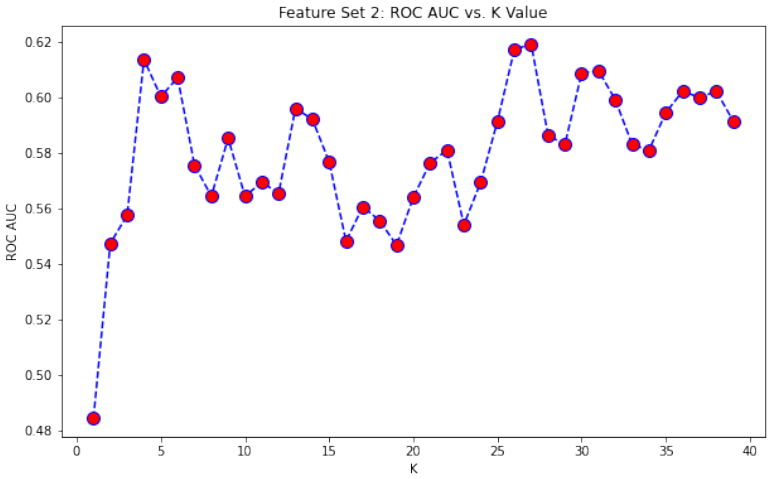
KNN classifier is a non-parametric classification algorithm. KNN makes use of proximity of groups around a data point to make classifications and predictions. For classifications, an input is labelled based on a plurality voting. This means that a data point is labelled the same as the most commonly group seen around that point. KNN classifiers are also known as "lazy" because they do not undergo a training process like traditional supervised learning algorithms, instead, they only store the training dataset. The computations are only made when testing data is used to make predictions. This can be a drawback as it can produce a high computation load when the model is deployed and needs to make a large number of predictions. This can make a KNN classifier difficult to scale. In order to make predictions, a number of distance metrics can be used by KNN classifiers. The most common ones used are Euclidean distance, Manhattan distance, Minkowski distance, and Hamming distance. The K value defined how many neighbors will be inspected to determine the label of a specific data point. Defining K is

an important part of the modeling as setting a K value that is too low can lead to high variance also known as overfitting and setting a K value that is too high can result in high bias also known as underfitting. The choice of K depends largely on the data. Data which has many outliers and high noise performs better with a higher K value [45], [20]. The distance metric for KNN was Euclidean distance and the number of neighbors were chosen by plotting the algorithms ROC curve AUC against the number of neighbors, picking the number for K that would maximise ROC curve AUC.

FIGURE 3.12: Depiction of how a KNN algorithm classifies new observations. [21]



(A) K-Value vs ROC AUC for BSNIP Feature Set 1



(B) K-Value vs ROC AUC for BSNIP Feature Set 2

FIGURE 3.13: K-Value vs ROC AUC for BSNIP Dataset by Feature Sets.

The values chosen for K by dataset and feature set are:

TABLE 3.12: Algorithm Parameters used for KNN by Dataset and Feature Set.

Dataset	Feature Set	K-Value
TBI	1	7
	2	5
BSNIP	1	22
	2	3

3.5.6 Oversampling

Oversampling is a resampling strategy applied when training a model on an imbalanced dataset. Oversampling allows for the redistribution of classes within a dataset. Oversampling supplements the training data with multiple copies of some of the minority classes. These multiple copies come from performing simple random sampling with replacement on the minority class in the training set.

For example, when working with the BSNIP dataset where in the training set with 345 subjects of which 136 are controls and 209 are cases, simple random sampling with replacement is performed on the controls so that there are 209 cases and 209 controls raising the total number of the training set to 418. This is depicted in table 3.13

TABLE 3.13: Example of oversampling in the BSNIP training dataset using "minority" strategy.

	Training Set (pre oversampling)	Training Set (post oversampling)
Control	136	209
Case	209	209
Total	345	418

Oversampling was applied only on the random forest and KNN algorithms, this decision was made due to the parametric many of the assumptions needed to use the parametric algorithms, especially logistic regression algorithm were violated, SVM algorithm also had similar performance when compared to logistic regression. The decision to not apply oversampling on the parametric methods (logistic regression and SVM) was made as oversampling would not do anything to affect the model assumptions and the results would still be unreliable. Additionally, oversampling was not applied on decision trees given that random forest are an ensemble

learning method that makes use of decision trees and bagging, meaning random forest would outperform perform decision trees, this result was also seen when oversampling was not applied.

3.6 Model Assessment:

The models' performance was assessed using a number of metrics. Firstly, a confusion matrix was made, and following that accuracy, precision, specificity, sensitivity, and f1 score were calculated. A receiver-operator curve (ROC) was also constructed and the area under the curve (AUC) was calculated. A prediction was classifier as a case if probability of bein a case was higher than 0.5.

The confusion matrix is a good tool to quickly investigate a model's performance. The confusion matrix provides important information such as the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). These values are categorized as such [19]:

		Predicted Label		
		Negative	Positive	Total
True Label	Negative	TN	$FP(typeIerror)$	$TN+FP$
	Positive	$FN(typeIIerror)$	TP	$FN+TP$
	Total	$TN+FN$	$FP+TP$	N

Performance metrics such as accuracy, precision, specificity, sensitivity and f1 score are measured using this confusion matrix.

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

$$Precision = \frac{TP}{TP+FP}$$

$$Specificity = \frac{TN}{FP+TN}$$

$$Sensitivity = \frac{TP}{TP+FN}$$

$$F1-score = 2*\left(\frac{Precision*Sensitivity}{Precision+Sensitivity}\right)$$

The receiver operating characteristic curve (ROC) is a graph which is made by plotting the true positive rate against the false positive rate at different discriminant threshold values. This graph shows the ability of the model to reliably classify binary output. The diagnostic ability of the model increases with the area under the curve (AUC) of an ROC curve. An AUC of 1 denotes a perfect classifier while and AUC of 0.5 denotes a random binary classifier (ex. Flipping a coin). A diagonal lines divides the graph, and a curve to the left side of the diagonal is favoured. [19]

Results

4.1 Prediction Performance Metrics

Table 4.1 provides model performance metrics for models trained on TBI dataset Feature Set 1. Included metrics are accuracy, precision, specificity, sensitivity, f1 score and the ROC curve AUC. It is seen in table 4.1 that many of the algorithms do not perform well on TBI Feature Set 1, especially the parametric models, logistic regression and SVM, this poor performance is seen as specificity is 0 and sensitivity is 1 meaning these models predicted all subjects as being a case. This does not allow for the model to provide any value. Similar performance is also seen in the random forest, KNN, and random forest trained on oversampled training set algorithms. Decision trees and KNN trained on oversampled training set perform best with and ROC AUC of 0.57 and 0.62 respectively. While these are the best performing algorithms on this dataset and feature set, their performance is lacking as an ROC AUC of 0.5 denotes a random classifier.

TABLE 4.1: Classification performance metrics on TBI dataset for algorithms using Feature Set 1.

Algorithm	Accuracy	Precision	Specificity	Sensitivity	F1 Score	ROC AUC
Logistic Regression	0.4783	0.4783	0	1	0.6471	0.47
SVM	0.4783	0.4783	0	1	0.6471	0.53
Decision Trees	0.5662	0.5333	0.4167	0.7173	0.6154	0.57
Random Forest	0.217	0.5	0.083	1	0.6667	0.74
KNN	0.217	0.5	0.083	1	0.6667	0.61
Oversampling + KNN	0.5652	0.5714	0.75	0.3636	0.4444	0.62
Oversampling + Random Forest	0.5652	0.5238	0.1667	1	0.6875	0.6

Table 4.2 provides performance metrics for models trained on TBI dataset using Feature Set 2. Similar trends are seen when compared to Table 4.1 in that the parametric algorithms do not perform well with a sensitivity of 1 and a specificity closer to 0. The non parametric algorithms perform the best with the best performance being from decision trees which has a high specificity and sensitivity as well and ROC AUC of 0.73. In the TBI dataset using Feature Set 2, oversampling did not help improve performance much.

TABLE 4.2: Classification performance metrics on TBI dataset for algorithms using Feature Set 2.

Algorithm	Accuracy	Precision	Specificity	Sensitivity	F1 Score	ROC AUC
Logistic Regression	0.5217	0.5	0.083	1	0.6667	0.39
SVM	0.5217	0.5	0.083	1	0.6667	0.54
Decision Trees	0.7826	0.875	0.9167	0.6363	0.7368	0.73
Random Forest	0.5652	0.5263	0.25	0.9091	0.6667	0.52
KNN	0.4348	0.4444	0.1667	0.7273	0.5217	0.58
Oversampling + KNN	0.5622	0.5656	0.6667	0.4545	0.5	0.55
Oversampling + Random Forest	0.5217	0.5	0.3333	0.7272	0.5926	0.56

Table 4.3 presents performance metrics for models trained on the BSNIP dataset using Feature Set 1. It is seen in this table that parametric models do not perform well with a specificity of 0 and a sensitivity of 1, meaning all subjects were classifier as a case. Better performance is seen when using oversampling to train the KNN and random forest models which have ROC AUC of 0.64 and 0.65 respectively.

TABLE 4.3: Classification performance metrics on BSNIP dataset for algorithms using Feature Set 1.

Algorithm	Accuracy	Precision	Specificity	Sensitivity	F1 Score	ROC AUC
Logistic Regression	0.6941	0.6941	0	1	0.8194	0.58
SVM	0.6941	0.6941	0	1	0.8194	0.60
Decision Trees	0.5882	0.7264	0.4423	0.6525	0.6952	0.56
Random Forest	0.6588	0.7272	0.3077	0.8136	0.67680	0.6
KNN	0.7	0.7557	0.3846	0.8390	0.7952	0.64
Oversampling + KNN	0.5941	0.7882	0.6538	0.5678	0.6600	0.64
Oversampling + Random Forest	0.5882	0.7449	0.5192	0.6186	0.6759	0.65

Table 4.4 presents performance metrics for models trained on the BSNIP dataset using Feature Set 2. It is seen in this table that parametric models do not perform well with a specificity of 0 and a sensitivity of 1, meaning all subjects were classifier as a case. This performance is also seen in the random forest algorithm. Better performance is seen when using decision trees with ROC AUC of 0.63 and when using oversampling to train the KNN and random forest models which have ROC AUC of 0.63 and 0.61 respectively.

TABLE 4.4: Classification performance metrics on BSNIP dataset for algorithms using Feature Set 2.

Algorithm	Accuracy	Precision	Specificity	Sensitivity	F1 Score	ROC AUC
Logistic Regression	0.6941	0.6941	0	1	0.8194	0.55
SVM	0.6941	0.6941	0	1	0.8194	0.40
Decision Trees	0.6352	0.7545	0.4808	0.7034	0.7281	0.63
Random Forest	0.6941	0.6941	0	1	0.8194	0.65
KNN	0.6412	0.7522	0.4615	0.7203	0.7359	0.62
Oversampling + KNN	0.5764	0.8026	0.7115	0.5169	0.6288	0.63
Oversampling + Random Forest	0.5705	0.8260	0.7692	0.4830	0.6096	61

4.2 Confusion Matrices:

Table 4.5 depicts the confusion matrices which were used to calculate the performance metrics in Table 4.1. It can be seen the parametric algorithms predict all subjects as a case, and random forests and KNN also have a similar issue, with the exception of 1 prediction. This is also seen to some extent when using random forest trained using oversampling.

TABLE 4.5: Confusion Matrices for algorithms using Feature Set 1 on TBI dataset

True Label \ Predicted Label	Negative	Positive
Negative	0	12
Positive	0	11

(A) Logistic Regression

True Label \ Predicted Label	Negative	Positive
Negative	0	12
Positive	0	11

(B) SVM

True Label \ Predicted Label	Negative	Positive
Negative	5	7
Positive	3	8

(C) Decision Trees

True Label \ Predicted Label	Negative	Positive
Negative	1	11
Positive	0	11

(D) Random Forest

True Label \ Predicted Label	Negative	Positive
Negative	1	11
Positive	0	11

(E) KNN

True Label \ Predicted Label	Negative	Positive
Negative	4	8
Positive	3	8

(F) Oversampling with KNN

True Label \ Predicted Label	Negative	Positive
Negative	2	10
Positive	0	11

(G) Oversampling with Random Forest

Table 4.6 depicts the confusion matrices which were used to calculate the performance metrics in Table 4.2. It can be seen the parametric algorithms predict almost all subjects as a case. Best performance is seen using decision trees which has a high number of true negatives and positives with an especially low number of false positives.

TABLE 4.6: Confusion Matrices for algorithms using Feature Set 2 on TBI dataset

		Predicted Label	
		Negative	Positive
True Label	Negative	1	11
	Positive	0	11

(A) Logistic Regression

		Predicted Label	
		Negative	Positive
True Label	Negative	1	11
	Positive	0	11

(B) SVM

		Predicted Label	
		Negative	Positive
True Label	Negative	11	1
	Positive	4	7

(C) Decision Trees

		Predicted Label	
		Negative	Positive
True Label	Negative	3	9
	Positive	1	10

(D) Random Forest

		Predicted Label	
		Negative	Positive
True Label	Negative	2	10
	Positive	3	8

(E) KNN

		Predicted Label	
		Negative	Positive
True Label	Negative	9	3
	Positive	7	4

(F) Oversampling with KNN

		Predicted Label	
		Negative	Positive
True Label	Negative	4	8
	Positive	3	8

(G) Oversampling with Random Forest

Table 4.7 depicts the confusion matrices which were used to calculate the performance metrics in Table 4.3. It can be seen the parametric algorithms predict all subjects as a case. Best performance is seen using KNN which shows a high number of true negatives and true positives and low numbers of false positives and negatives.

TABLE 4.7: Confusion Matrices for algorithms using Feature Set 1 on BSNIP dataset

		Predicted Label	
		Negative	Positive
True Label	Negative	0	52
	Positive	0	118

(A) Logistic Regression

		Predicted Label	
		Negative	Positive
True Label	Negative	0	52
	Positive	0	118

(B) SVM

		Predicted Label	
		Negative	Positive
True Label	Negative	23	29
	Positive	41	77

(C) Decision Trees

		Predicted Label	
		Negative	Positive
True Label	Negative	16	36
	Positive	22	96

(D) Random Forest

		Predicted Label	
		Negative	Positive
True Label	Negative	20	32
	Positive	19	99

(E) KNN

		Predicted Label	
		Negative	Positive
True Label	Negative	34	18
	Positive	51	67

(F) Oversampling with KNN

		Predicted Label	
		Negative	Positive
True Label	Negative	27	25
	Positive	45	73

(G) Oversampling with Random Forest

Table 4.8 depicts the confusion matrices which were used to calculate the performance metrics in Table 4.4. It can be seen the parametric algorithms predict all subjects as a case, this is also seen for the random forest model. Better performances are seen using random forest and KNN trained using oversampling.

TABLE 4.8: Confusion Matrices for algorithms using Feature Set 2 on BSNIP dataset

		Predicted Label	
		Negative	Positive
True Label	Negative	0	52
	Positive	0	118

(A) Logistic Regression

		Predicted Label	
		Negative	Positive
True Label	Negative	0	52
	Positive	0	118

(B) SVM

		Predicted Label	
		Negative	Positive
True Label	Negative	25	27
	Positive	35	83

(C) Decision Trees

		Predicted Label	
		Negative	Positive
True Label	Negative	0	52
	Positive	0	118

(D) Random Forest

		Predicted Label	
		Negative	Positive
True Label	Negative	24	28
	Positive	33	85

(E) KNN

		Predicted Label	
		Negative	Positive
True Label	Negative	37	15
	Positive	57	61

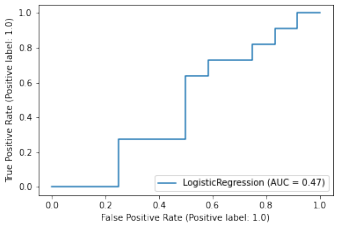
(F) Oversampling with KNN

		Predicted Label	
		Negative	Positive
True Label	Negative	40	12
	Positive	61	57

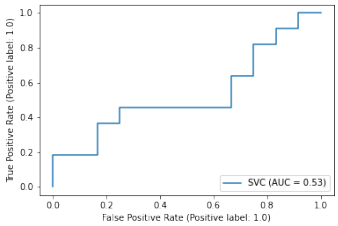
(G) Oversampling with Random Forest

4.3 ROC Curves:

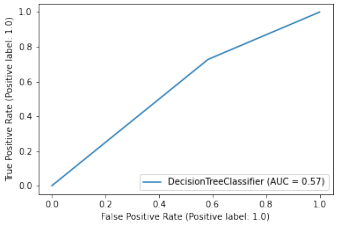
Figure 4.1 depicts ROC curves for the models trained on the TBI dataset using Feature Set 1. The highest ROC AUC is 0.74 for random forests (subplot D). While the AUC is high the model performance is poor as specificity is close to 0 and sensitivity is 1.



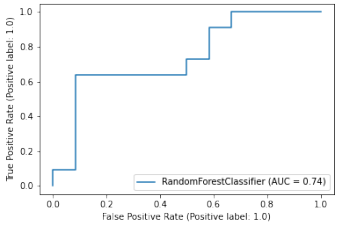
(A) ROC curve for logistic regression



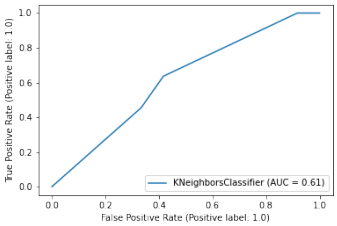
(B) ROC curve for SVM



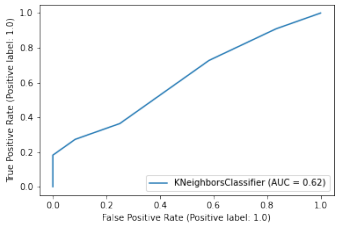
(C) ROC curve for decision trees



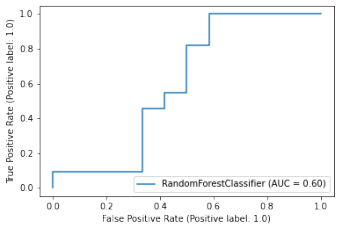
(D) ROC curve for random forests



(E) ROC curve for KNN



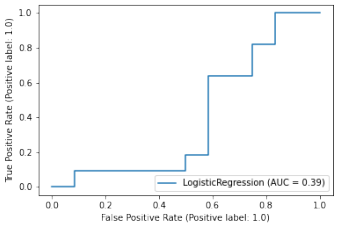
(F) ROC curve for oversampling with KNN



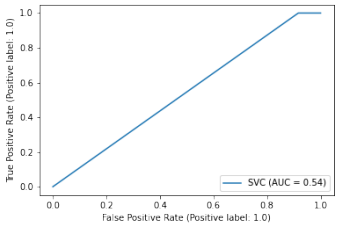
(G) ROC curve for oversampling with Random Forests

FIGURE 4.1: ROC curves for algorithms using Feature Set 1 on TBI dataset

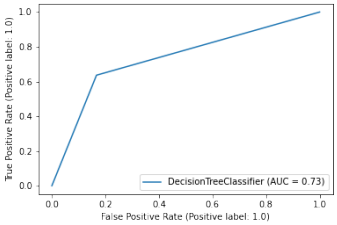
Figure 4.2 depicts ROC curves for the models trained on the TBI dataset using Feature Set 2. The highest ROC AUC is 0.73 for decision trees (subplot C). This model performs well and predicted low amounts of false negatives and false positives.



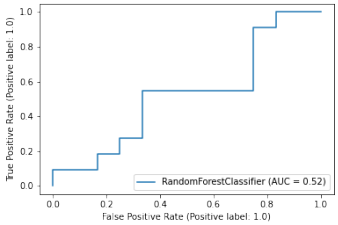
(A) ROC curve for logistic regression



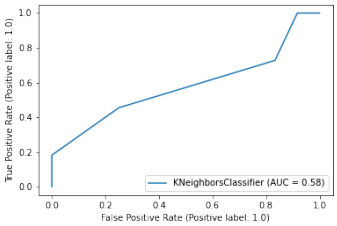
(B) ROC curve for SVM



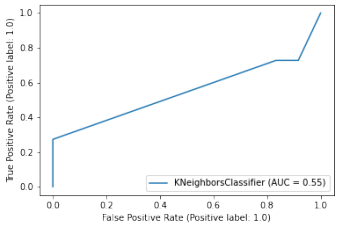
(C) ROC curve for decision trees



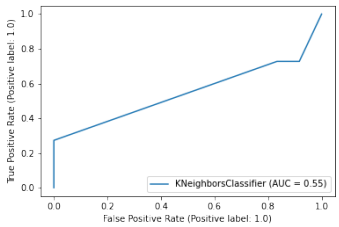
(D) ROC curve for random forests



(E) ROC curve for KNN



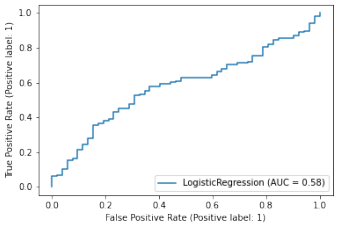
(F) ROC curve for oversampling with KNN



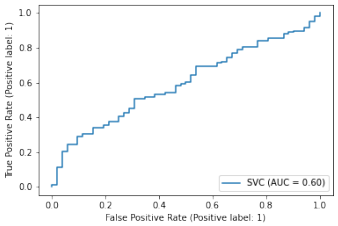
(G) ROC curve for oversampling with Random Forests

FIGURE 4.2: ROC curves for algorithms using Feature Set 2 on TBI dataset

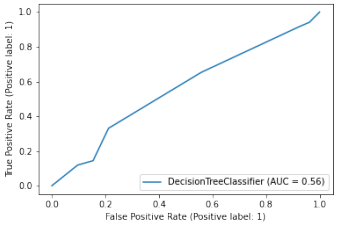
Figure 4.3 depicts ROC curves for the models trained on the BSNIP dataset using Feature Set 1. The highest ROC AUC is 0.65, seen for random forest trained using oversampling (subplot G). This model performs well and predicted low amounts of false negatives and false positives which relatively high levels of specificity and sensitivity.



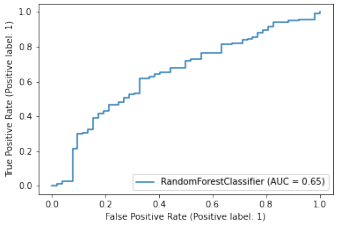
(A) ROC curve for logistic regression



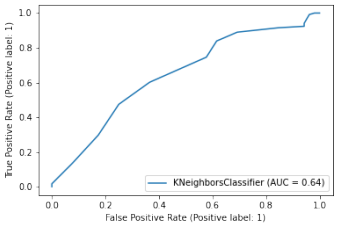
(B) ROC curve for SVM



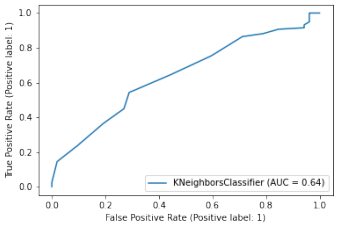
(C) ROC curve for decision trees



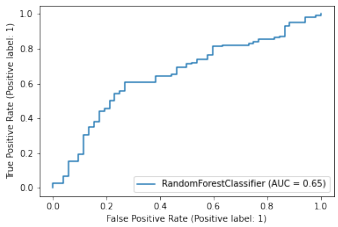
(D) ROC curve for random forests



(E) ROC curve for KNN



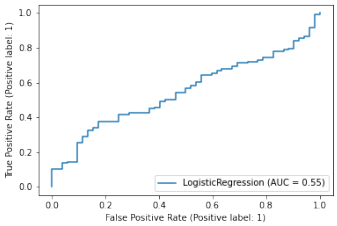
(F) ROC curve for oversampling with KNN



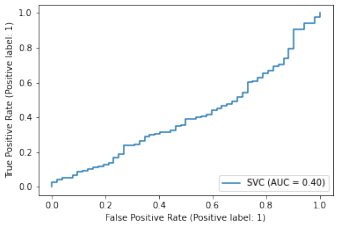
(G) ROC curve for oversampling with Random Forests

FIGURE 4.3: ROC curves for algorithms using Feature Set 1 on BSNIP dataset

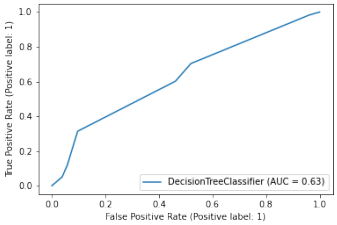
Figure 4.4 depicts ROC curves for the models trained on the BSNIP dataset using Feature Set 2. The highest ROC AUC is 0.65, seen for random forest(subplot D). This model does not perform well as it has a specificity of 0 and sensitivity of 1.



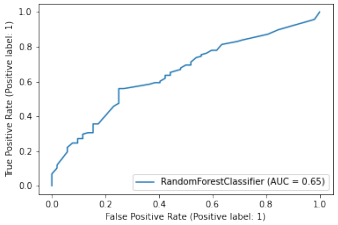
(A) ROC curve for logistic regression



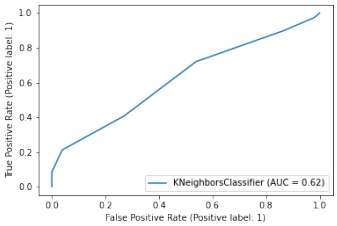
(B) ROC curve for SVM



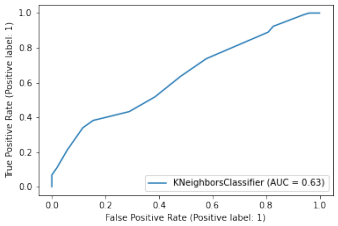
(C) ROC curve for decision trees



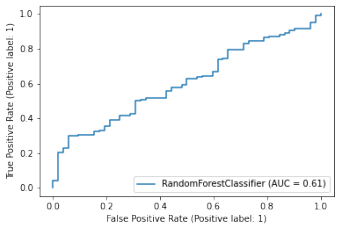
(D) ROC curve for random forests



(E) ROC curve for KNN



(F) ROC curve for oversampling with KNN



(G) ROC curve for oversampling with Random Forests

FIGURE 4.4: ROC curves for algorithms using Feature Set 2 on BSNIP dataset

In conclusion, in the TBI dataset, regardless of feature set, decision trees performs the best. It can also be said that the parametric approaches do not provide a useful model. Feature Set 2 provides more useful models (models not predicting all subjects as case) when compared to Feature Set 1.

Similar conclusions can be drawn when using the BSNIP dataset. Parametric approaches do not provide a useful model (predict all subjects as cases) using Feature Set 1 or Feature Set 2. Random forests trained using oversampled training set perform best when using Feature Set 1. KNN (regular and trained on oversampled training set), and random forest trained using oversampled training set perform best when using Feature Set 2.

Discussion

When sourcing the data, some criteria were set:

1. The dataset should have a large number of subjects as some data-hungry algorithms such as SVM, and random forests were to be used.

2. The dataset should have subjects with multiple different mental illnesses and controls in it as the study developed EEG features for general mental health status as opposed to features for the diagnosis of specific illnesses.

3. The dataset should contain EEG reading data and mental health status data for an auditory oddball experiment with multiple trials. The oddball experiment was a requirement as the 2 ERP components of interest in this study, the P300 and N200, are reliably produced in an auditory oddball experiment and multiple trials were required to establish an ERP.

The dataset we found that best fit all these criteria was the Bipolar and Schizophrenia Network on Intermediate Phenotypes (BSNIP) dataset available (after some delay) through the National Institute of Mental Health Data Archive (NDA) website [39]. In addition to the BSNIP dataset, a traumatic brain injury (TBI) and healthy control dataset found on OpenNeuro was also chosen [7]. This TBI dataset failed to meet the criterion of multiple mental illnesses and only contained data for TBI and healthy controls but it was immediately available and allowed data analysis methods to be tried out early on while still awaiting BSNIP access.

As mentioned above, one of the main ways the TBI and BSNIP data differ from each other is that the TBI dataset only contains data for one mental health issue (TBI) while the BSNIP dataset contains data for subjects who have multiple mental illnesses (schizophrenia, schizoaffective disorder, and bipolar disorder, and psychosis) [39]. This means that in the TBI dataset, the feature set cannot be tested for predicting general mental health status as there is only one

mental illness in that dataset, the results on this dataset are only for predicting TBI. In order to determine the efficacy of the features for general mental health status we tested on the BSNIP dataset.

To give more details on the use of two datasets, we note the two datasets were chosen as the TBI dataset was easier to work with due to its smaller size (fewer subjects), low complexity (one mental illness and healthy controls), and took less computing power during feature extraction. The BSNIP scripts took about 2 hours to run while TBI took 15 minutes. Also the data use approval process from the NIMH for the BSNIP dataset was lengthy and the outcome of the process was uncertain, therefore it was decided to do feature development on the TBI dataset in parallel with the data use approval process for the BSNIP dataset. This decision was made to reduce the overall time needed to complete the thesis

This choice to develop the features on the TBI dataset first has had some unintended effects on the nature of the features. The TBI dataset had a low subject count compared to the BSNIP dataset. One of the ways it changed the feature development process is that due to the low subject count, the feature set had to be low in dimensions as having a high dimensional feature set would result in overfitting (ie. poor generalization). Given that the TBI dataset only had 48 subjects in its training set and using the rule of thumb of 5 observations per parameter [24], this would only allow for about 10 parameters for a feature set.

Benefits of being restricted in this way are that one is forced to only develop features that are efficient in classifying mental health status and increasing the interpretability of the input features. A possible drawback of being limited in dimensionality is that due to the low dimensionality of the input feature set, the model may have high bias due to there not being enough features to train a model that captures the true complexity of the data (ie. Underfitting).

After the datasets were selected, while inspecting the datasets, it was decided to use binary outcome variables for this study. For the TBI dataset, the outcome variable is "HadHeadInjury" which denotes whether the subject has had a TBI or not, and for the BSNIP dataset, the variable "Phenotype/diagnosis for the subject" which denoted whether the subject was a case or a control. A benefit of using these variables was that these were based off of assessments made by physicians as opposed to relying on symptom ratings or an analog measurement. A drawback of using binary variables was that information regarding illness severity is lost and further analysis on the prediction of severe illness vs mild illness cannot be conducted.

A consideration to make is that for the BSNIP dataset, only the mental illness status (ie. case vs control) of the subject was available. Specificity about the diagnosis for each subject was not available. This data would have been valuable as it would allow for the stratification of results to see if there is a correlation between the predictive ability of the features and the

mental illness the subject has. This would have further allowed for this study to test the true generalizability of the feature sets.

A concern with 74% of the original TBI dataset being used after data cleaning is the introduction of attrition bias in the remaining data. If there is a systematic difference between the subjects that are included and not included in the final data this could lead to under coverage. In the case of this dataset, the number of subjects were reduced due to information on their mental illness status was not available, and further information as to why was not available. Another concern was the imbalance between the TBI subjects and the controls as well as the low number of the control subjects. Some implication of this imbalance is high prediction error [40]. This may show up, in extreme cases, as models that predict all subjects as being cases, which is not useful. A way to combat this is either redistributing the classes through resampling or adjusting the weights of the models, since these two methods have a similar result, in this study oversampling was used.

Similarly, in the BSNIP dataset, which started off with 613 unique subjects with EEG readings available, 515 remained. In this dataset, the relative loss of subjects was lower which means a relatively lower attrition/selection bias in the BNSIP dataset. In the case of the BSNIP dataset the loss was due to the subjects not having the EEG channels required to calculate the P300 and N200 ERPs. This could be caused due to random error, such as improper connections, channels being detached, or faulty equipment, but the true source of this error is unknown. While the relative loss of subjects is lower, the imbalance between cases and controls is almost the same between both datasets. One benefit of the BSNIP dataset is that due to the higher total subject count in the BSNIP dataset, data sparsity is less of a concern. Oversampling is also used in the BSNIP dataset in an effort to fix the imbalance between the controls and cases.

When developing the features, as seen in figures 3.5 the values of P300 peak amplitudes and latencies differ depending on the channel. If the peak amplitude and latencies of each channel per ERP component were used as a feature set, this would result in a feature set with 24 features, which given the low number of subjects in the TBI dataset is too many dimensions for the feature set to have enough predictive ability as too many dimensions would lead to sparsity of data resulting in unreliable predictions.

As an alternative, the average and variance of the peak amplitudes and latencies across the 6 channels for P300 and N200 ERP components were calculated. This resulted in a feature set with 8 features which would satisfy the rule of thumb of 5 observations per parameter [24]. The goal was to describe P300 and N200 peaks in as few dimensions as possible as it would estimate the mean amplitudes and latencies, and the variance value would still provide information on the distribution of the different amplitudes and latencies produced by multiple channels. This resulted in some loss of information when averaging multiple values. A benefit of averaging was that the signal to noise ratio was increased, resulting in features that would reflect the true

value of the ERP components. A possible detriment of using this feature set is that it may not capture enough data needed to make an accurate and reliable prediction. By describing ERP components by just 2 values, a lot of valuable information is lost, such as multiple peaks as seen in figures 3.5 and 3.6, differing shapes of the peaks, different widths of the peaks, possible plateauing at the peaks value as seen in figure 3.5 a etc. This could lead to underfitting (high bias).

The second feature set that was developed was with the same goal of describing the P300 and N200 ERP components in mind. This time, it is done by calculating the mean amplitude over the P300 and N200 time window. For the P300, the mean amplitude was taken over ms 250-400 after the stimulus, while for N200 the mean amplitude was taken over ms 200-350 after the stimulus. This was done for each channel and then the values were averaged and the variance was calculated over the 6 centro-parietal and fronto-central channels respectively. This resulted in a 1X4 feature set. One key manner this feature set is different from the first feature set is that data regarding P300 and N200 latency is missing as the mean amplitude is calculated across the whole ERP component window. This feature set has 2 benefits compared to the first feature set, one, it has lower dimensionality, meaning that the data is less sparse, and two, the mean amplitude acts as a low pass filter to increase the signal to noise ratio of the ERP component, giving a more accurate reading of the true ERP component value. A possible disadvantage of this feature set is the same as Feature Set 1 that a lot of possibly valuable data, especially to do with ERP component latency, is lost which can lead to an increase in bias(ie. underfitting).

Conclusion

Several studies have made use of EEG features to detect specific mental health illnesses but general mental health diagnostic tools (biomarker or symptom-based) to identify individuals who are manifesting early signs of mental health disorders are not commonly available. This thesis seeks to explore the potential use of EEG features as a biomarker-based tool for general mental health diagnosis [8].

Two feature sets were developed and tested in this study. The study was conducted using 2 datasets. 5 algorithms were used to test the predictive ability of the feature sets, Logistic regression, support vector machines, decision trees, random forests, KNN classification algorithms were used, additionally random forest and KNN were also trained using oversampling to improve performance. The model performance was tested using accuracy, precision, sensitivity, specificity, f1 score, confusion matrices, and ROC AUC. In general non parametric model performed better than parametric models. Feature Set 1 had better performance when using the BSNIP dataset while Feature Set 2 had better performance when using the TBI dataset. The best performing algorithms across the board were random forest and KNN both trained using oversampling, and decision trees.

These results show promise in the use of EEG features to predict general mental health and additionally show that there may be shared EEG characteristics in multiple mental illnesses. This study's results invite further research in this area. The use of this technology opens doors for biomarker-based diagnosis of mental health conditions, lowering the cost of mental health care, and making mental health care accessible for more people [23].

Bibliography

[1] Asieh Ahani, Helane Wahbeh, Hooman Nezamfar, Meghan Miller, Deniz Erdogmus, and Barry Oken. Quantitative change of eeg and respiration signals during mindfulness meditation. *Journal of neuroengineering and rehabilitation*, 11(1):1–11, 2014.

[2] Betul Ay, Ozal Yildirim, Muhammed Talo, Ulas Baran Baloglu, Galip Aydin, Subha D Puthankattil, and U Rajendra Acharya. Automated depression detection using deep representation and sequence learning with eeg signals. *Journal of medical systems*, 43:1–12, 2019.

[3] Francisco Barcelo. Detection of change: event-related potential and fmri findings: John polich (ed.); kluwer academic publishers, boston, usa, 2003, hardcover, 187 pp., isbn 1-4020-7393-3, 2004.

[4] Vladimir Bostanov, Lilian Ohlrogge, Rita Britz, Martin Hautzinger, and Boris Kotchoubey. Measuring mindfulness: a psychophysiological approach. *Frontiers in human neuroscience*, 12:249, 2018.

[5] James Broadway, Rebecca Rieger, Richard Campbell, Davin Quinn, Andrew Mayer, Ronald Yeo, J. Wilson, Darbi Gill, Violet Fratzke, and James Cavanagh.

[6] James Cavanagh, Rebecca Rieger, J. Wilson, Darbi Gill, Lynne Fullerton, Emma Brandt, and Andrew Mayer. Joint analysis of frontal theta synchrony and white matter following mild traumatic brain injury. 2020.

[7] James F Cavanagh and Davin Quinn. "eeg: Three-stim auditory oddball and rest in acute and chronic tbi". 2021.

[8] James F Cavanagh, J Kevin Wilson, Rebecca E Rieger, Darbi Gill, James M Broadway, Jacqueline Hope Story Remer, Violet Fratzke, Andrew R Mayer, and Davin K Quinn. Erps predict symptomatic distress and recovery in sub-acute mild traumatic brain injury. *Neuropsychologia*, 132:107125, 2019.

[9] Bruno A Cayoun. *Mindfulness-integrated CBT: Principles and practice*. John Wiley & Sons, 2011.

[10] Michael Chmielewski, Lee Anna Clark, R Michael Bagby, and David Watson. Method matters: Understanding diagnostic reliability in dsm-iv and dsm-5. *Journal of abnormal psychology*, 124(3):764, 2015.

[11] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[12] E Jane Costello. Early detection and prevention of mental health problems: developmental epidemiology and systems of support. *Journal of Clinical Child & Adolescent Psychology*, 45(6):710–717, 2016.

[13] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. In *Machine learning techniques for multimedia*, pages 21–49. Springer, 2008.

[14] Brittany F Escuriex and Elise E Labbé. Health care providers' mindfulness and treatment outcomes: A critical review of the research literature. *Mindfulness*, 2(4):242–253, 2011.

[15] Bradley D Grinage. Diagnosis and management of post-traumatic stress disorder. *American Family Physician*, 68(12):2401–2408, 2003.

[16] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[17] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[18] Stefan G Hofmann, Alice T Sawyer, Ashley A Witt, and Diana Oh. The effect of mindfulness-based therapy on anxiety and depression: A meta-analytic review. *Journal of consulting and clinical psychology*, 78(2):169, 2010.

[19] Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1, 2015.

[20] Qinghua Hu, Daren Yu, and Zongxia Xie. Neighborhood classifiers. *Expert systems with applications*, 34(2):866–876, 2008.

[21] IBM. What is the k-nearest neighbors algorithm? `https://www.ibm.com/topics/knn`, 2023. [Online; accessed 11-January-2023].

[22] Bogumił Kamiński, Michał Jakubczyk, and Przemysław Szufel. A framework for sensitivity analysis of decision trees. *Central European journal of operations research*, 26(1):135–159, 2018.

[23] Eran Klein. Ethics and the emergence of brain-computer interface medicine. *Handbook of clinical neurology*, 168:329–339, 2020.

[24] Konstantinos Koutroumbas and Sergios Theodoridis. *Pattern recognition*. Academic Press, 2008.

[25] Jeffrey A Lieberman, Scott A Small, and Ragy R Girgis. Early detection and preventive intervention in schizophrenia: from fantasy to reality. *American Journal of Psychiatry*, 176(10):794–810, 2019.

[26] Kim-Lian Lim, Philip Jacobs, Arto Ohinmaa, Donald Schopflocher, and Carolyn S Dewa. A new population-based measure of the economic burden of mental illness in canada. *Chronic Dis Can*, 28(3):92–98, 2008.

[27] Andreas Müller, Sarah Vetsch, Ilia Pershin, Gian Candrian, Gian-Marco Baschera, Juri D Kropotov, Johannes Kasper, Hossam Abdel Rehim, and Dominique Eich. Eeg/erp-based biomarker/neuroalgorithms in adults with adhd: Development, reliability, and application in clinical practice. *The World Journal of Biological Psychiatry*, 2019.

[28] Peter E Nathan and Jack M Gorman. *A guide to treatments that work*. Oxford University Press, 2015.

[29] Tanya Navaneelan. Suicide rates: An overview. 2012.

[30] Abraham M Nussbaum et al. *The Pocket Guide to the DSM-5-TR™ Diagnostic Exam*. American Psychiatric Pub, 2022.

[31] Payam Refaeilzadeh, Lei Tang, and Huan Liu. Cross-validation. *Encyclopedia of database systems*, 5:532–538, 2009.

[32] Darrel A Regier, Emily A Kuhl, and David J Kupfer. The dsm-5: Classification and criteria changes. *World psychiatry*, 12(2):92–98, 2013.

[33] Juan Ruiz de Miras et al. Schizophrenia classification using machine learning on resting state eeg signal. 2022.

[34] Sidney J Segalowitz and Kerry L Barnes. The reliability of erp components in the auditory oddball paradigm. *Psychophysiology*, 30(5):451–459, 1993.

[35] Paul Smetanin, Carla Briante, Minhal Khan, David Stiff, and Sheeba Ahmad. The life and economic impact of major mental illnesses in canada. 2015.

[36] Jill C Stoltzfus. Logistic regression: a brief primer. *Academic emergency medicine*, 18(10):1099–1104, 2011.

[37] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974.

[38] Shravani Sur and Vinod Kumar Sinha. Event-related potential: An overview. *Industrial psychiatry journal*, 18(1):70, 2009.

[39] Carol A Tamminga, Godfrey Pearlson, Matcheri Keshavan, John Sweeney, Brett Clementz, and Gunvant Thaker. Bipolar and schizophrenia network for intermediate phenotypes: outcomes across the psychosis continuum. *Schizophrenia bulletin*, 40(Suppl_2):S131–S137, 2014.

[40] Erdal Tasci, Ying Zhuge, Kevin Camphausen, and Andra V Krauze. Bias and class imbalance in oncologic data—towards inclusive and transferrable ai in large scale oncology data sets. *Cancers*, 14(12):2897, 2022.

[41] Michael E Thase. Atypical depression: useful concept, but it's time to revise the dsm-iv criteria. *Neuropsychopharmacology*, 34(13):2633–2641, 2009.

[42] Daniel Vigo, Graham Thornicroft, and Rifat Atun. Estimating the true global burden of mental illness. *The Lancet Psychiatry*, 3(2):171–178, 2016.

[43] JC Wakefield. Dsm-5, psychiatric epidemiology and the false positives problem. *Epidemiology and Psychiatric Sciences*, 24(3):188–196, 2015.

[44] Wikipedia. Support vector machine — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Support%20vector%20machine&oldid=1129233078`, 2023. [Online; accessed 11-January-2023].

[45] Wenchao Xing and Yilin Bei. Medical health big data classification based on knn classification algorithm. *IEEE Access*, 8:28808–28819, 2020.