Supporting Materials

Callum Carville

List of abbreviations

BBC	British Broadcasting Corporation
NHS	National Health Service
CAGR	Compound Annual Growth Rate
WHO	World Health Organisation
GHDx	Global Health Data Exchange
SMI	Severe Mental Illness
COPD	Chronic Obstructive Pulmonary Disease
CHD	Coronary Heart Disease
POC	Proof of Concept
SMD	Severe Mental Disorders
fMRI	functional Magnetic Resonance Imaging
MDD	Major Depressive Disorder
DSM-IV	Diagnostic and Statistical Manual-IV
DSM	Diagnostic and Statistical Manual
EEG	Electroencephalogram
NN	Neural Networks
RNN	Recurrent Neural Networks
CNN	Convolutional Neural Networks
LSTM NN	Long-Short Term Memory Neural Networks
BCI	Brain Computer Interface
CRISP-ML	Cross-Industry Standard Process for Machine Learning
TDSP	Team Data Science Process
CRISP-DM	Cross-Industry Standard Process for Data Mining
PTSD	Posttraumatic Stress Disorder
OCD	Obsessive Compulsive Disorder
IDE	Integrated Development Environment
FFT	Fast Fournier Transformation
OVR	One Versus Rest
PSD	Power Spectral Density
DL	Deep Learning
ML	Machine Learning
UL	Unsupervised Learning
SSL	Semi Supervised Learning
SL	Supervised Learning
IBM	International Business Machines
SVM	Support Vector Machine
OVR	One vs Rest
RF	Random Forest
API	Application Programming Interface
SMOTE	Synthetic Minority Oversampling Technique

Initial aims and objectives set

1. Determine the best research method for the project.
2. Retrieving a non-biased EEG data with accurate labels for EEG signals.
3. Determine if the data requires alterations to be suitable for processing
4. Determine the current state of the art algorithms for processing time series data (EEG is time series based).
5. Discuss strengths & weaknesses of particular algorithms for classification.
6. Develop a model that can be submitted as part of a POC.
7. Develop a model that can achieve a 95% minimum accuracy to meet a professional standard.
8. Develop a model that can achieve generalisation.

Context and Literature Investigation

Number of individuals taking antidepressants

BBC news article by Annabel Rackham [1]

According to the NHS the number of individuals taking antidepressants has increased by 5% during the period 2021-2022, this marks the sixth year in a row that there has been an increase in both patients and prescriptions. An estimated 83.4 million antidepressant drug items were prescribed to 8.3 million patients (this was based on UK residents). These statistics indicate the need for an improvement in the diagnostic and treatment process of those suffering from Depression or related illnesses as treatments should reflect in less patients not more.

Antidepressants Global Market

Report by The Business Research Company [2]

The antidepressants market grew from an estimated $16.44 billion in 2022 to $17.41 billion in 2023 with a CAGR of roughly 5.9%. This market size and growth rate demonstrates an alarming need to improve the treatment of patients and find cures versus more customers which is reflected in the trend of increased market size.

Mental Disorders

Article by WHO [3]

An article published by the WHO with supporting data from GHDx [4] made the bold state 1 in 8 people during 2019 in the world lived with a mental disorder (these ranged from emotional regulation, behaviour disorders & disturbances in thinking). This article made huge headlines globally and reflected a need for a solution to improve mental health amongst the population.

Psychiatrists Global Market

Report by Research and Markets [5]

Psychiatrists provide the diagnosis, treatment, and prevention of emotional, mental, and behavioural disorders. Psychiatrists focus on curing the following schizophrenia, depression or anxiety-based disorders. The market has grown in this field in 2022 from $147.28 billion to $164.01 billion in 2023 meaning a growth CAGR of 11.4%. These stats along with those provided by the NHS and Business Research Company show an alarming onboarding of additional patients instead of a decrease indicating that the processes are failing. The need for a change is now otherwise we are draining resources of the economy, reducing the number of individuals fit to work and most importantly not solving the disorders which serves as a moral duty.

Premature mortality in adults with severe mental illness (SMI)

Report by Office for Health Improvement & Disparities [6]

This report describes how people with SMI (this includes schizophrenia, bipolar disorder, depression & other psychotic illnesses) suffer from premature mortality and it can vary by gender, age, socioeconomic group, and geographical area in England. The latest study focuses on data acquired from 2018 to 2020 which states that there is a correlation between SMI and poor physical health. Those suffering with SMI often develops chronic conditions such as obesity, asthma, diabetes, COPD, CHD, stroke, heart failure and liver disease at a younger age than people without SMI.

The physical health problems are linked to risk of premature death in those suffering with SMI, but SMI is often not recorded as an underlying cause of death or recorded on death certificates as a contributory cause which is a surprising fact given the strong correlation between the two. The study estimates that 2 out of 3 people with SMI who die from a physical illness that could have been prevented. The study consisted of 3 years' worth of data with 120,273 (70,533 males and 49,740 females) adults with SMI between the ages of 18 to 74. In conclusion the study found that adults suffering from SMI are 2.5 to 7.2 times more likely to die prematurely (before the age of 75) than adults who do not have SMI in England.

Mortality of people with severe mental illness: Causes and ways of its reduction.

Published on National Library of Medicine by Mario Luciano, Maurizio Pompili, Norman Sartorius, and Andrea Fiorillo [7]

In this 2022 publication the researchers found that SMDs are heavily associated with a variety of other illnesses, and those suffering tend to have poorer health outcomes and in effect a higher rate of mortality than those with other non-communicable diseases. The publishers state that individuals suffering with SMD die on average 10-20 years earlier than the general population and this is accompanied by an increase in standardized mortality rates. This additional publication is used to reinforce the severity and potential dangers to those suffering from psychiatric disorders, further justifying the need improved diagnosis and treatment.

How self-esteem relates to physical & mental health

Article posted 2022 by Cognition coaching & consulting. [8]

The article states that self-esteem is directly related to physical and mental health with low self-esteem leading to poor health behaviours such as smoking, alcohol abuse, obesity, depression, and drug abuse. It was even stated that low self-esteem is one of the main risk factors for physical illness.

Arxiv.org Paper

Depression Diagnosis & Drug Response Prediction via Recurrent Neural Network and Transformers Utilizing EEG Signals Paper by Abdolkarim Saeedi, Arash Maghsoudi, Fereidoun Nowshiravan Rahatabad, Department of biomedical engineering, Science and research branch, Islamic Azad university, Tehran, Iran, Department of Medicine, Baylor College of Medicine, Houston, TX, United States Center for Innovations in Quality, Effectiveness, and Safety, Michael E. DeBakey VA Medical Center, Houston. [9]

The paper released in 2023 discusses how the early diagnosis and treatment of depression is vital for effective treatment. The paper criticises the current understanding of the disease in both research and clinical practice. The paper proposes a method to diagnose a patient with the disorder as well as a method to predict the drug response in a patient using EEG signals.

The publishers of this paper made use of Transformers (modified RNN [10] with an alternative architecture) to evaluate the time series data and compared this model against CNN [11], LSTM and CNN-LSTM [12]models. The Transformer model achieved an average recall of 99.41% with an accuracy of 97.14% for classifying either a healthy or depressed patient.

The results from this study reinforce the novelty of Transformers replacing recursive models as a new structure to examine EEG data [13] and retain a high level of accuracy. This paper serves as the foundation for investigating Transformer models to classify other psychiatric conditions via EEG [13] recordings.

Prediction of Depression from EEG Signal Using Long Short-Term Memory (LSTM)

Publication by S.Dhananjay Kumar & DP Subha [14]

This publication utilises LSTM models [15] to predict Depression instants via EEG recordings, the study consists of 30 patients at resting state and extracts 5600 samples to train the model. The root mean square error for the model was 0.000064% showing success in the experiment and effectively proving EEG can be used to detect Depression patterns.

Machine learning identification of EEG features predicting working memory performance in schizophrenia and healthy adult

Publication by Jason K Johannsesen, Jinbo Bi, Ruhua Jiang, Joshua G Kenney and Ch-Ming A Chen [16]

This study consisted of 40 diagnosed with Schizophrenia patients along with 12 healthy controls whom all completed a working memory task (Sternberg Working Memory Task) whilst being recorded with EEG. SVM [17] was the classifier choice for this research and the team were able to successfully distinguish subjects suffering from Schizophrenia with 87% accuracy.

The patients diagnosed with Schizophrenia were required to meet the DSM-IV criteria before being eligible to take part in the study. Subjects were also included if they met the age requirements which were marked for subjects to be at least 18 years old, be native English speaking and have a stable housing situation for a minimum of 30days. Patients diagnosed with the disorder were excluded if their medication had changes within the last 30 days or if they abused alcohol or drugs, had previous brain trauma or suffered from mental retardation. Further interviews were conducted to obtain treatment, substance use, medical history and psychosocial background information.

To extract the EEG recordings all participants were placed in front of a 24" monitor at a viewing distance of 1metre in a dimly lit room. EEG was recorded with a 64-channel BioSemi ActiveTwo bio-amplifier with electrodes placed according to the 10-20 system. Extra electrodes were placed at mastoids to act as a reference point, on the outer of both eyes, above and below the right orbit. EEG was acquired with a sampling rate of 1024Hz per second with a notch filter of 60Hz.

When applying the SVM classifier the team mentioned there was a risk of overfitting as they had extracted 60 EEG features which outweighed the number of sample classes to work from. This also meant that training could achieve excellent score accuracy but poor validation/testing accuracy. To combat this the team applied 1-norm regularisation (loss function). When the features were processed via SVM the classifier reportedly only utilised 3-10 features making more than 89% of the data redundant which is an interesting statistic in itself. The C parameter of the SVM model was tuned with the a 3-fold cross-validation process which split the data intro a training of 66.66% and testing of 33.33%. SVM model identified gamma activity during the encoding and occipital theta as the main features for their research.

This study demonstrates how ML architectures along with EEG can be used for feature detection in a binary classification scenario. The study also concluded features derived match the literature which is that gamma's role in memory encoding and theta's role during memory retention along with supported recordings for elevated resting low frequency activity in patients diagnosed with Schizophrenia and that SVM is a suitable ML method for classification between patients suffering with the disorder versus healthy controls.

Identification of major psychiatric disorders from resting-state electroencephalography using a machine learning approach

Publication by Su Mi Park, Boram Jeong, Da Young Oh, Chi-Hyun Choi, Hee Yeon Jung, Jun-Young Lee, Donghwan Lee, Jung-Seok Choi and sponsored by Department of Psychiatry, SMG-SNU Boramae Medical Center, Seoul, South Korea, Department of Statistics, Ewha Womans University, Seoul, South Korea, Department of Psychiatry and Behavioral Science, Seoul National University College of Medicine, Seoul, South Korea and Institute of Human Behavioral Medicine, Seoul National University Medical Research Center, Seoul, South Korea. [18]

This publication not only is the foundation of the project but it was also responsible for developing the dataset that was selected to be used. The aim of this project was the develop a ML classifier to detect and compare psychiatric disorders using EEG. SVM, Random Forest and Elastic net were the architectures used for the classification of the data and the paper concludes that ML in

EEG could be used to predict psychiatric disorders and make distinctions between a healthy control and those with a disorder. This is extremely promising research that shows ML could have a potential future in EEG psychiatric disorder predictions and diagnosis.

Elastic net along with the IQ adjustment performed the best and the team also recorded the ideal feature combinations which consisted of for Schizophrenia the best features were PSD alpha with a 93.83% accuracy, for Trauma and Stress-related disorders the best feature was beta FC achieving a 91.21%, for Anxiety disorders the whole band of PSD reflected the best features with the accuracy of 91.03%, for Mood disorders theta FC was the best feature combination achieving 89.26% accuracy, Addictive disorders had theta PSD as the best features with a 85.66% accuracy and for OCD disorder gamma FC were the best features achieving 74.52% accuracy.

The research conducted to acquire the dataset was obtained through a psychiatrist based on the DSM-IV or DSM-5 criteria along with using the Mini-International Neuropsychiatric interview during psychological assessments. The final diagnosis was confirmed by 2 psychiatrists and 2 psychologist and their decision came from a review of the original patient diagnosis with the psychological assessments which were completed 1 month before and post EEG recording. All patients were over the age of 18 meaning the EEG recordings were of both adult male and female brains (this helps with training data) and individuals were excluded if the patient suffered a brain injury or it their IQ was sub <70 which is the metric used to determine intellectual function.

The study concludes that more severe disorders such as Schizophrenia and PTSD were easier to discriminate against and thus achieve a high classification accuracy score. The team did state that this may be a factor of both disorders being associated with altered brain activity. Theirs findings also mention how neurodynamical state heterogeneity has the possibility to exists with 2 or more disorders such as when attempting to classify Mood disorder versus Depression disorder as they share overlapping symptoms. The team also shared their opinion on how this research might be useful for diagnostic decisions between psychiatric disorder and the research could be further extended by attempting multi-class classification methods to determine the usability of EEG ML.

Limitations of the study included the medications the patients were taking and the level of how severe the disorder was not controlled. Patients who were recorded by the EEG may have been suffering from additional disorders which were not distinctly identified when recording. The study was of a local nation and not a global study meaning recordings from various other nationalities may skew results and as a result generalisation is not possible with the dataset used.

EEG Source Network for the Diagnosis of Schizophrenia and the Identification of Subtypes Based on Symptom Severity—A Machine Learning Approach

Article by Jeong-Young Kim, Hyun Seo Lee and Seung-Hwan Lee [19]

This article published 04/12/2020 performs research to determine if it possible via EEG to classify Schizophrenia subtypes based on the severity of the symptoms. The study was conducted with 64 electrodes applied to 119 patients diagnosed with Schizophrenia (a mix of 53 males and 66 females) with 199 healthy controls (a mix of 51 males and 68 females) during resting-state. The patients with the disorder were divided into 2 groups high and low via the PANSS. The classification accuracy was evaluated via 10-fold cross-validation with the LDA classifier. The team achieved an accuracy of 80.66% for binary classification for patients with the Schizophrenia disorder and healthy controls along with also accumulating accuracy scores for distinguishing between low and high categories in positive, negative and cognitive symptoms with a peak of 88.10%, 75.25% and 77.78% accuracy.

The article focuses on the previous diagnostic criteria from the ‘Diagnostic and Statistical Manual of Mental Disorders’ / DSM-5 by interviewing the patients with a series of in-depth questions designed to extract information including the duration of the disorder and the presence of the symptoms. These questions are flawed as patients may lie about their symptoms and the symptoms provided by the patients overlap with other disorders making it difficult to truly with 100% confidence distinguish a patient with Schizophrenia. To combat these difficulties researchers resorted to other means for their diagnosis via using tools such as EEG as a cost-effective solution for obtaining biomarkers.

The literature mentions how through EEG and fMRI disruptions were found in several cortical regions which included the prefrontal, parietal and temporal lobes of the brain. Decreased connectivity between the cerebellum and the prefrontal cortex were noted to correlate with increased negative symptoms of Schizophrenia whilst positives symptoms displayed a strong correlation between functional connectivity in the posterior cingulate and middle temporal regions of the brain.

Patients’ psychiatric symptoms were evaluated with the PANSS methodology to determine which were fit for the study. The results lead to the exclusion of patients with life time history of central nervous system disease, alcohol/drug abuse, mental retardation and head injuries resulting in loss of consciousness. This left 119 patients of which 25 were drug naïve and 94 were taking antipsychotic medications of the following prescriptions; aripiprazole: 11, amisulpride: 10, blonanserin: 6, clozapine: 5, haloperidol: 1, olanzapine: 16, paliperidone: 11, quetiapine: 10, risperidone: 22, ziprasidone: 1 and zotepine: 1.

For the EEG recordings patients were placed on a chair with ambient noise blocked, the patients’ eyes were closed for 4 minutes and EEG signals were recorded with the Quick Cap with 62 electrodes which were applied to the subject’s scalp in accordance to the 10-20 system. The data was recorded at a sampling rate of 1000Hz per second and removed 60Hz noise with the notch filter. An individual inspector with acclaimed ‘training’ identified and manually removed artifacts found in the recordings which consisted of eye blinks and eye movement. For feature classification sequential forward selection was applied which is a bottom up searching technique that selects the best features first according to a cost function. The frontal and parietal lobes were frequently identified as the best features for classification accuracies.

The teams research concluded that resting-state EEG could successfully perform binary classification between Schizophrenia patients and healthy controls along with identifying the differentials between low and high categories when performing classification.

A machine learning framework involving EEG-based functional connectivity to diagnose major depressive disorder (MDD)

Article by Wajid Mumtaz, Syed Saad Azhar Ali, Mohd Azhar Mohd Yasin and Aamir Saeed Malik [18]

This publication focuses on MDD otherwise referred to as Depression classed as a mental illness/psychiatric disorder and makes use of binary classification between patients diagnosed with MDD versus healthy controls with the following algorithms, SVM, Logistic Regression and Naïve Bayes.

The teams study proved a success as they achieved accuracy scores well above the realm of chance with SVM performing the best with an accuracy score of 98%, Logistic Regression reinforced the success of the team with 91.7% accuracy and lastly Naïve Bayes achieved a score of 93.6% accuracy. They concluded by stating SL could be a promising method for diagnosing patients which could become a generalised approach with further tool developments.

The researchers have mentioned with the ML techniques reflecting in promising results they are still yet to recognised in clinical translation. The research was performed on a total of 64 subjects, 34 with Depression (18 females and 16 males) and 30 healthy controls (9 females and 21 males). The subjects with the Depression disorder were recruited from an outpatient clinic hospital called Universiti Sains Malaysia and the patients had to meet the diagnostic criteria for DSM-IV. Healthy subjects were also examined for clinical symptoms to help exclude the possibility of any mental disability which could skew results. All participants in the study were instructed to abstain from coffee, nicotine and alcohol before the EEG recordings and were recorded at the same time of the day to keep consistency across recordings.

The EEG recordings consisted of 5-minute rest-stating with eyes closed and eye open as the 2 separate conditions. The data was acquired through 19 electrodes covering the scalp with placements being determined by the 10-20 electrode placement standards. A 50Hz notch filter was applied and a sample rate of 256 per second was used to record the electrical activity of the subject's brain. The EEG recordings were processed to remove artifacts that could skew results, these artifacts may include blinking, muscle activations or even the subjects heart beating. The reduced multiple source modelling technique was noted as the method to remove noise from the recordings as part of preprocessing the data.

The team used 10-fold cross validation to evaluate classification performance which meant the training testing split was 90% for training and 10% for testing which is above the suggested practice and opinions on ResearchGate however this was a small dataset meaning data was limited. 3 metrics were evaluated accuracy, sensitivity and specificity, for my project the accuracy from these results was the most vital as it determined whether or not classification of Depression was even possible with EEG.

The publication concluded for the study SVM classifier with the linear kernel performed the best over nonlinear polynomial and the Gaussian kernel with the belief it is because the sample size for training data was limited making it insufficient for variations of the SVM classifier. The team believe the results from their research make ML with EEG feasible for clinical applications as they can discriminate between the subject with the disorder and the healthy control with more than just chance. They also propose an automated diagnosis system that can be utilised to assist in clinics. However, with the study being on a small sample size can make it difficult to determine its ability to generalise if the system was commercialised.

Deep learning applied to electroencephalogram data in mental disorders: A systematic review

Article by Mateo de Bardeci, Cheng Teng lp & Sebastian Olbrich [18]

This article makes observations on how DL architectures have been applied EEG for both diagnostic and predictive use for mental/psychiatric disorders. The study views in particular how CNN or LSTMs networks were used for classification with EEG studies on psychiatric diseases based on ICD-10 or DSM-v. The team made observations on 3 main categories, clinical, EEG data processing and DL architectures used by other researchers. The team made the bold statement that the acquisition and pre-processing of EEG signals was sufficient in many studies but many of them lacked systematic characterisation of clinical features and many models were inappropriately used with flawed testing metrics.

The clinical category consisted of observing the studies information with regards the description of the diagnosis validation, the number of subjects included for with the number of which the sample extracted, the information about the subjects' medical treatments and clinical use for the paper (whether it was a diagnostic approach, disorder prediction or drug response prediction.

EEG-data category contained the conditions under which the EEG was recorded (e.g resting state, sleep etc…), the sampling rate of the recording (the number of EEG signals stored per second), the number of EEG channels being recorded and how the artefacts were removed.

DL category was observed for information about how the studies performed feature engineering, network structure, optimisation algorithms, model architecture selection and testing of the model.

The team first began by querying the database for relevant papers published in English until the date 21.10.2020 with the search phrases set to 'electrophysiological measure/EEG', 'DL networks and terms for mental disorders that were set according to F-code diagnosis of the ICoD. The results after eliminating those not relevant consisted of 30 published papers in this field. An assigned author extracted then a list of variables from the 3 categories above to highlight the areas to review in the papers. Next papers were excluded if they contained any references to either Dementia or Parkinsons disease as they are listed as neurological conditions in ICD-10 and studies involving the improvement of sleep were also removed.

The publications were ranked by the level of reproducibility of the DL with the levels being set to ‘poor’, ’middle’ or ‘good’. A paper would be rate ‘poor’ if essential aspects were missing making the DL model impossible to reproduce, ‘middle’ if all essentials’ aspects were accessible with the possibility of reproducibility but still not specific enough and ‘good’ is for when all essentials are present, reproducibility is possible and the paper is specific.

The study concludes that emerging analysis techniques in EEG research can enhance how researchers utilise clinical EEG and how DL models can support the extraction of patterns from EEG time series which in the past may have been hidden within complex structures. The team also provided 11 suggestions for researchers to improve DL models with EEG studies and these are; using clear terminology, be precise when describing clinical samples with identification of cofounding variables, validate the diagnostic procedures against international standards, follow EEG standards for recording and processing, explore data augmentation, select a clear model strategy and make sure to test the model, ensure that test data and training data are independent, make sure to identify and balance cofounding variables, select appropriate scoring measures for reporting such as F1-score, analyse and report the influence of hyperparameters and lastly improve transparency via in depth descriptions of the models and make code publicly available on repositories such as GitHub. This publication will influence my EEG project as there have been many key points raised to help improve the research conducted and the criteria vital in producing a respected report.

Business Understanding

Why this research is important?

With a rising rate in the number of people suffering from psychiatric disorders along with the increase of the pharmaceutical market for prescriptions relating to these disorders, it calls for an improvement in the diagnostic and analysis process of these diseases. With supporting evidence from the research paper found on Arxiv.org [9] stating that there is a lack of clinical understanding for certain illnesses such as depression.

There can be an ethical question raised regarding the correlation of increased individuals suffering from psychiatric disorders and the growth of the pharmaceutical market who are responsible for developing drugs to treat people suffering, not profiting from their disorder. If the treatments were effective and cured the patient, surely a decrease in the number of individuals suffering should be the result not the opposite?

With a better understanding of the disorder and feedback from drugs prescribed recorded, a better treatment or cure can be implemented. This should reduce the number of people suffering, improve their quality of life and improve our societies understanding of these disorders.

In the more extreme cases people suffering from these disorders tend to also suffer from a poor physical health which in turn can take its toll on the individual’s self-esteem and in worst cases lead to a fatality. This is the motive behind developing better diagnostic models than can help classify a range of psychiatric conditions which can then be used to help provide more accurate treatment to the patient.

When researching this space there was a lack of available articles, publications or reports that focus on developing models with the up to date or current popular architectures which include LSTM NN [15] and Transformer based models. This research not only help with the classification of psychiatric disorders based on EEG data [19] but also strengthen the availability of publications and research in the field.

Is there any commercial value in the research?

The psychiatric pharmaceutical market has grown from $147.28 billion USD to $164.01 billion USD in 2023, [5] this upward trend in the market creates a need for a better diagnosis and treatment for the patient. Creating generalised models that can help classify a patient’s psychiatric disorder [20] with high level accuracy will help identify the disorder faster and therefore an appropriate treatment can be provided.

Another benefit for creating generalised models is the improvement of the treatment for the patient as the models can be used as a metric comparison to record if the treatment is having a positive, neutral, or negative effect on the disorder. The models can help streamline the development of new drugs or cures as less time and resources would need spent on the diagnostic stage of the process.

How this research should be approached?

To begin this research selecting a well labelled and publicly available dataset will be necessary to develop a proficient model that can then be used to convey how as a POC it is possible to classify a patient from a range of available psychiatric disorders that EEG data [19] has been recorded for.

The published paper which studies the diagnosis of depression with drug response prediction via recurrent neural networks and transformers utilising EEG signals [13] [21] will be used to form a structure of my own research as this paper shares a lot of similarities in terms of techniques used and the data type selected for my project. This paper [9] also demonstrates how the use of recurrent neural networks and long-short term memory networks can develop models with over 95% accuracy in both classification of an individual with depression and recall.

The paper starts with an introduction, identification of previous work, the materials required for the research, the selected methods to develop the models, testing the models and then the evaluation of the models along with graphs which demonstrate the training, evaluation curves of the models developed, a discussion and finally a references section for sourced materials. Mimicking this papers format will help to develop the optimal paper in line with other publications standards in this field.

With my style of project, I have selected to perform correlation research as the focus of this methodology is to find patterns and relationships in the EEG data [19] deriving facts with probabilities to determine which category best fits a recorded individual. This type of research approach best suits my data as the dataset contains multiple variables including the individuals age, gender, IQ, time series EEG [19] recordings along with patient diagnosis.

When reviewing the type of research being undertaken, I found it shared more similarities with qualitative research as the primary aim is to explore the concept of applying alternative ML [22]architectures and algorithms to develop models on EEG data [19] to classify disorders. The data selected is also from a study group of individuals already diagnosed with a psychiatric disorder making it possible to confirm a model's ability to handle unseen data which can help a model become generalised.

Life cycle stages

The stages of this project's life cycle shared the most similarity to the CRISP-ML [23] process model. The reason being is that the project majorly focused on developing a ML model which is a subset field within machine learning whereas CRISP DM [24] [25] focuses primarily on data mining and analytics.

CRISP-ML methodology consists of seven stages:

Business Understanding – Within this stage the problem or opportunity is identified using machine learning techniques and the projects objectives and success criteria would be set.

Data Understanding – Data is collected and observed to improve understanding whilst determining relevance to the problem set in the prior stage. The data is observed to determine if there are quality issues or limitations which would need addressed to make the dataset suitable for the set task.

Data Preparation – With a dataset selected, preprocessing is used to clean the data, handle the missing values, outliers with other inconsistencies.

Modelling – DL algorithms are selected based on the datatype and problem type which in this case is a classification problem. The data is split into training, validation, and testing samples (this is important for the evaluation step). The model is then developed with the training data and optimised by fine tuning the models hyperparameters to achieve better performance whilst avoiding overfitting or underfitting. Models are compared with one another, and their performance and accuracy are used as the measurements to determine which model was the best performing.

Evaluation – The model's performance is evaluated against the test set and the success of the model will be determined on the objectives previously set which is above a 95% accuracy in the classification of a psychiatric disorder.

Deployment – With the model defined it would next be integrated into a production environment with the appropriate infrastructure and software.

Maintenance – The model would then be monitored and updated to ensure the performance remains optimal and achieve a high level of accuracy.

Technologies used

Python is the selected programming language that will be used to develop the models as it is well supported with external libraries to help assist in development. [26]

PyCharm was the IDE used to write Python scripts in that were not compatible with the Jupyter environment and allow for visuals to be developed of the electrode placement. [27]

Excel was used to view the EEG data before loading into Python to get a better understanding of both the format and data types. [28]

CSV file format was used to load the EEG data into the Python environment and be converted to a data frame. [29]

Pandas is a software library developed for Python and used for data manipulation. I used this plugin to convert CSV based data to a data frame for processing. [30]

NumPy is a Python library that adds support for multi-dimensional arrays, matrices and a series of mathematical functions for handling arrays. This was particular useful when handling the EEG data as large dimensions were used during computation. [31]

Matplotlib is a Python library that added additional support for embedding plots into the Jupyter notebook environment which was particularly useful for data visualisation and comparing results from various algorithms. [32]

Seaborn is similar to Matplotlib except is allows for additional high level data visualisation plots and interfaces which helped improve the overall data understanding. [33]

Cat Boost is an algorithm based on gradient boosting with decision trees. This algorithm was used with several others in attempt to perform multi class classification tasks and help determine the disorder of each patient based on EEG recordings. [34]

Sklearn KNeighborsClassifier is an algorithm based on selecting data points need a K value. This algorithm was used with several others in attempt to perform multi class classification tasks and help determine the disorder of each patient based on EEG recordings. [35]

Lightgbm LGBMClassifier is an algorithm based on constructing a gradient boosting model. This algorithm was used with several others in attempt to perform multi class classification tasks and help determine the disorder of each patient based on EEG recordings. [36]

Sklearn linear_model is an algorithm that fits a linear model with coefficients/weights/features to minimize the residual sum of squares between the observed targets in the dataset. This algorithm was used with several others in attempt to perform multi class classification tasks and help determine the disorder of each patient based on EEG recordings. [37]

Sklearn.naive_bayes GuassianNB uses the probability of each class by calculating the mean and standard deviation for the training data. This algorithm was used with several others in attempt to perform multi class classification tasks and help determine the disorder of each patient based on EEG recordings. [38]

Sklearn.ensemble RandomForestClassifier is an algorithm that fits several decision tree classifiers on various samples of the data whilst using averaging to improve the overall accuracy. This algorithm was used with several others in attempt to perform multi class classification tasks and help determine the disorder of each patient based on EEG recordings. [39]

Sklearn tree is an algorithm based on a tree structure. This algorithm was used with several others in attempt to perform multi class classification tasks and help determine the disorder of each patient based on EEG recordings. [40]

Xgboost XGBClassifier is an algorithm that utilises gradient boosted decision trees to improve overall speed and performance. This algorithm was used with several others in attempt to perform multi class classification tasks and help determine the disorder of each patient based on EEG recordings. [41]

Sklearn RepeatedKFold was used for cross validation of the data with the repeated K fold method set to 5 splits meaning 80% of the data was used for training and 20% was used for testing. [42]

Matplotlib pyplot is a library for python to develop visualisations and in this it was used to develop boxplots to observe the variance of results from the models processed with cross validation. [43]

Adobe Photoshop is a visual editing tool which aided in the development of a diagram that explained the data flow of the project to complement the methodology used. [44]

Data Understanding and Data Preparation

What is EEG? [13] [21] [19] [45] [46]

An EEG is a recording that measures electrical activity of the brain using electrodes (metals discs) along with gel to improve conductivity to the scalp. The signals are produced by neurons which are the main cells of the human nervous system. Neurons are comprised of the following components:

Dendrites - Extending from the cell body are dendrites which cover a large surface area for receiving signals.

Nucleus – Responsible for cell function and regulation.

Axons – The axon is important as it is the conduction region of the neuron as it generates electrical signals for communicating with other neurons. This electrical signal also referred to as impulse it passed to the terminus to trigger the release of neurotransmitters.

Axon hillock - Extruding from the cell body is the axon hillock which acts as the connection between a cell body and axon.

Myelin sheath - The axons are coated/protected in myelin sheath which increase the speed at which impulses are transmitted.

Schwann cells – These individual units make up the myelin sheath with their main task to envelop the nerve Fiber, wrapping around it several times to create the plasma membrane that protects the nerve.

Node of Ranvier – Between each Schwann cell there is an area of Fiber exposed where axon collaterals branch out.

Terminus – Thousands of terminal branches which interface with other neurons.

Glial cells - Delicate parts of the neuron are wrapped in glial cells for protection.

There are several types of neurons which can be classed by structure or function, according to Professor Dave the main neurons by structure are unipolar, bipolar, and multipolar. These naming conventions refer to the number of processes extending from the cell body.

A unipolar neuron contains a single axon which divides into proximal and distal branches, one of these is a central process that goes towards the nervous system and the other is a peripheral process that acts as a sensory receptor.

Bipolar neurons have two processes one axon with one dendrite which extend from opposite sides of the cell body.

Multipolar neurons have three or more processes with one being an axon and the rest are dendrites (an estimated 99% of a human's neurons are this type).

When classifying neurons by function there are three main types:

Sensory/Afferent neurons – Transmits data from sensory receptors towards central nervous system (commonly unipolar).

Motor/Efferent neurons – Transmits information from the central nervous system to muscles/glands (always multipolar).

Interneurons – Reside between and help pass signals around the central nervous system (commonly multipolar).

How does a neuron produce an electrical impulse? [13] [21] [19] [45] [46]

Opposite charges attract meaning for those charges to separate work must be done, if they are separated the opportunity to use their charge separation energy is available. The charge separation exists within nerve cells due to the concentrations of specific ions found inside and outside of the plasma membrane. This means there is a potential difference across the cell membrane.

The cell membrane resists current flow as formally charged ions have difficulty trans versing the nonpolar section of the membrane. To combat this ion channels located within the membrane let specific ions through at specific times. These are separated into non-gated (remains open) and gated (closed until a signal is received).

Chemically gated channels open when a specific neurotransmitter binds.

Voltage gated channels respond to the changes in membrane potential.

Mechanically gated channels open when the receptor is physically deformed.

Ions freely diffuse through when the above-mentioned channels are opened, these ions obey the electrochemical gradient which sole purpose is to balance charge, so an electrical signal is generated.

A resting neuron has a resting membrane potential (resting membrane potential is -70mV) which is based on the sodium and potassium ion concentrations with in and out of the cell due to their different ability to diffuse throughout the cell, these levels are maintained by sodium-potassium pumps that maintain the concentration gradient. A resting neuron will contain more sodium ions outside than in and more potassium ions inside than outside.

When a signal is received there will be a change in this potential which will either be graded potential (operates over short distances) or an action potential (operates over long distances e.g., length of axon).

In the case of an action potential depolarisation (lasts about 0.5ms) must exceed a threshold (typically around -55mV), this means sodium channels must open with enough sodium ions into the cell, if achieved a nerve impulse result with a current propagating along the axon towards the axon terminal. Repolarisation (lasts about 0.5ms) will occur next; this is where voltage-gated potassium channels allow potassium ions to exit the cell. This exit causes a short burst hyperpolarization (lasts about 2.5ms) before resetting to original levels.

Once the impulse meets the axon terminal releasing neurotransmitters into the synaptic space. A synapse is the junction between two neurons and in this space, communication takes place through a series of synapses.

A pre-synaptic neuron releases neurotransmitters that are generated through the cytoplasm at each of the thousands of axon terminals to activate receptors on the post-synaptic neuron triggering an action potential that propagates the signal to the corresponding neuron at a rate of an estimated one hundred metres per second.

This whole process begins with ligand-gated channels which are activated by neurotransmitters that end up producing a graded potential causing voltage gated channels to open and if enough ions diffuse that surpass the threshold the action potential is generated.

Once all these functions are completed a variety of membrane proteins will allow ion concentrations to reset including sodium-potassium pumps which make use of active transport to shuttle sodium and potassium ions back to their original sides of the membrane restoring the concentration gradient that provide resting potential. This leads to the absolute refractory period which lasts between one and two milliseconds where the neuron is incapable of firing again and finalising with the neuron in a resting state. This serves as important biological function as it ensures signals travel in one specific direction along an axon whilst preventing the neuron repeatedly firing (the estimated maximum a neuron can fire is 1000 times per second).

Selecting Appropriate Public Data

EEG data recordings of multiple variants of psychiatric disorders with diagnosis is in very limited supply to the public. However, I found one dataset that covered a range of psychiatric disorders [47] with other potential important features including: age, gender, age, date of EEG recording, education, IQ, disorder, specified disorder, and EEG site recordings. The data also varied in terms of the individuals being studied as there was a range of males and females with the age being as low as 18 and the upper age bracket including 70-year-olds. The data consisted of 945 subjects with 850 patients diagnosed with major psychiatric disorders and 95 healthy control patients. The EEG signals [13] were retrieved from a rest-state, and this was consistent throughout.

With the dataset being developed and published by Su Mi Park [48] an employee of ORCiD (organisation for connecting research and researchers) [49] along with the Department of Psychiatry in Borame Medical Centre (Seoul, Republic of Korea), the Department of Statistics in Ewha Womans University (Seoul, Republic of Korea), Department of Psychiatry and Behavioural Science in National University College of Medicine (Seoul, Republic of Korea) and the Institute of Human Behavioural Medicine in National University Medical Research Centre (Seoul, Republic of Korea) I can assume the data recorded can be trusted as there were many critical research observants involved.

Unfortunately, the raw EEG [13] recordings could not be sourced, I contacted the researchers who developed the dataset and was not successful in getting a response. As a result, I was forced to work with the pre-processed data which does not include the time series EEG recordings for the channels only the extracted means of each channel per band. This eliminated potential architectures such as LSTM, Transformers or CNNs. [15] [10] [11] [12]

Data Processing and Cleaning

The dataset contained fields which were not marked to act as features / weights [50] in the development of this classifier as I only required the EEG recordings along with the diagnosis of the disorder. The reason being is that as part of the objectives the aim was to determine if it is possible with EEG recordings [13] to classify an individual with the correct disorder, this way the model's ability to become more dynamic and generalised to the public is possible.

Prior to collecting this dataset, the research team who developed it also performed pre-processing of the EEG [13] signals and removed artifacts from the data. Artifacts are signals recorded by the EEG equipment [51] that are not generated from the brain, these artifacts [52] can skew results when utilising the data, so removal is very important.

With irrelevant data removed from the dataset, a check was performed for null values (values that are missing from the dataset which could impact results) which returned 0 for missing entries, this keeps the data supplied to the model consistent as there are no missing values. (The dataset used for visualisation was modified to include a separator column to separate the columns of data that contained the specified disorder along with their band and channel).

The diagnosis field contained data relating to disorder of the individual that was recorded, I modified the naming conventions of the disorders to remove white spaces in the data as IDE [53] compilers often have difficulty handling spaces in the data, when replacing the titles, I used a camel case approach to the naming conventions to match professional software engineering practices.

The next step in the data preparation process was to convert the text-based diagnosis field results into a binary format (converting a string to an integer), the reason for this being is that machine learning algorithms cannot be applied to string-based data, the data needs converted to a numerical value to be processed.

Finally, there was a check for outliers to determine if there were any discrepancies in the diagnosis of the individuals or instances where there is little supporting data as models are data driven. There were no outliers found within the dataset. With more reliable data fed into the model, the more it can learn and therefore improve the accuracy of results when diagnosing individuals.

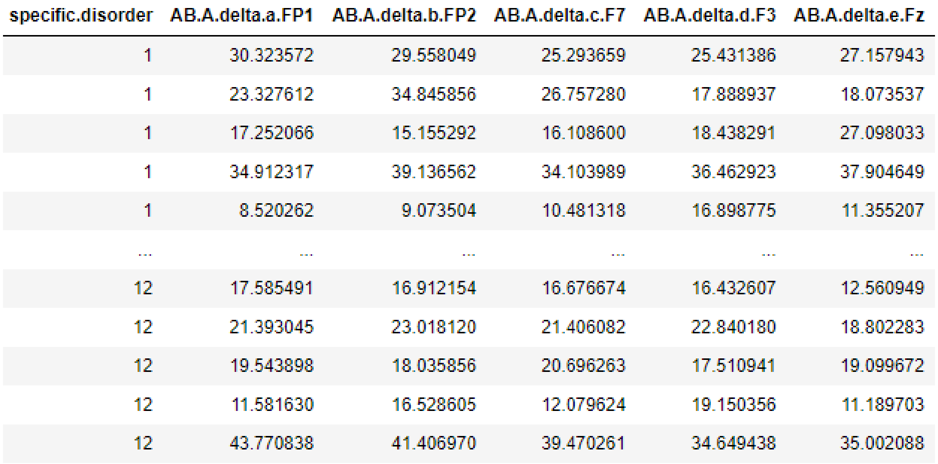
Data Visualization and Understanding

The resulting processed dataset left 946 rows with 1142 columns with 1 column dedicated to the disorder diagnosis, 114 Power Spectral Density (PSD) [54] EEG columns (19 electrodes * 6 frequency bands) and 1026 columns representing coherence of EEG (measured between every pair of electrodes for each frequency band making the overall calculation 171 * 6).

The PSD [54] columns provide a method for representing the distribution of an EEG signal [13] frequency making them more interpretable and their values represent the Watts/Hz at which signals were recorded. Each PSD [54] measures the signal of power contributed by frequencies within a band.

Coherence measures the synchronisations between signals of two different electrodes and is based on phase consistency. Frequency and phase consistency are derived from the EEG time series data [13].

The screenshot below shows a limited example of the data along with the column headings, all column headings that begin with ‘AB’ represent PSD and the columns starting with ‘COH’ represent coherence. Odd numbers in the headings represent the left side of the head and the even numbers the right side. The capital letters in front of the numbers at the end of the heading names stand for F for front, P for parietal, T for temporal and O for occipital for each lobe of the brain.



specific.disorder	AB.A.delta.a.FP1	AB.A.delta.b.FP2	AB.A.delta.c.F7	AB.A.delta.d.F3	AB.A.delta.e.Fz
1	30.323572	29.558049	25.293659	25.431386	27.157943
1	23.327612	34.845856	26.757280	17.888937	18.073537
1	17.252066	15.155292	16.108600	18.438291	27.098033
1	34.912317	39.136562	34.103989	36.462923	37.904649
1	8.520262	9.073504	10.481318	16.898775	11.355207
...	...	...	...	...	...
12	17.585491	16.912154	16.676674	16.432607	12.560949
12	21.393045	23.018120	21.406082	22.840180	18.802283
12	19.543898	18.035856	20.696263	17.510941	19.099672
12	11.581630	16.528605	12.079624	19.150356	11.189703
12	43.770838	41.406970	39.470261	34.649438	35.002088

Figure 1 - Snippet to preview cleaned EEG Dataset layout

The diagnosis column contained a healthy control (95 participants) along with EEG [13] recordings for individuals with PTSD [55] (52 participants), Schizophrenia [56] (117 participants), Depression [57] (119 participants), Social anxiety disorder [58] (48 participants), Bipolar disorder [59] (67 participants), OCD [60] (46 participants), Alcohol use disorder [61] (93 participants), panic disorder [62] (59 participants), adjustment disorder [63] (38 participants), behavioural disorder [64] (93 participants), and acute stress disorder [65] (38 participants).

PTSD [55] – This is a disorder that develops in individuals who have experienced a life shocking or dangerous event.

Schizophrenia [56] – This mental illness affects an individual’s feelings, thoughts, and behaviours.

Depression [57] – This disorder ranges from long lasting feelings of unhappiness, hopelessness to loss of interests in the things that you enjoyed or gave you meaning.

Social anxiety disorder [58] – With this disorder an individual suffers from a fear that does not alleviate with every day social activities and directly negatively affects self-confidence.

Bipolar disorder [59] – This mental illness can either be chronic or episodic with extreme mood swings which when severe can cause impairment in social or occupational functioning.

OCD [60] – Individuals suffering with OCD have uncontrollable or reoccurring thoughts and or behaviours that they feel they need to repeat over and over.

Alcohol use disorder [61] – An individual suffering with this condition has an impaired / inability to stop or control their alcohol use despite the health or social consequences.

Panic disorder [62] – This disorder is where an individual suffers from regular recurring panic attacks which impacts their ability to process thoughts or act rationally.

Adjustment disorder [63] – This is when an individual suffers from an emotional response to stress related events.

Behavioural disorder [64] – This diagnosis applies when an individual's behaviours are disruptive, persistent, or severe.

Acute stress disorder [65] – This mental health issue stems / occurs in the first month post traumatic event and shares symptoms like PTSD [55].

Healthy control – This reflects mind of a healthy individual who has not been diagnosed by a professional with any disorder and was used as a comparison control group for the experiment.

Under and over sampling [66] [66] was also used to combat the imbalanced data as an alternative approach for model development. Under sampling is when examples from the training data set that belong to a majority class are removed in order to improve the balance of the overall data set. The 'RandomUnderSampler' [66] method from imbalanced learn API was used to perform under sampling and balance the training data. The screen shot below displays the quantity of each disorder post processing leaving a much smaller dataset that is balanced. This API randomly selects samples to remove with the option for replacement, in this instance I did not implement replacement as I wanted to use as much real data as possible.

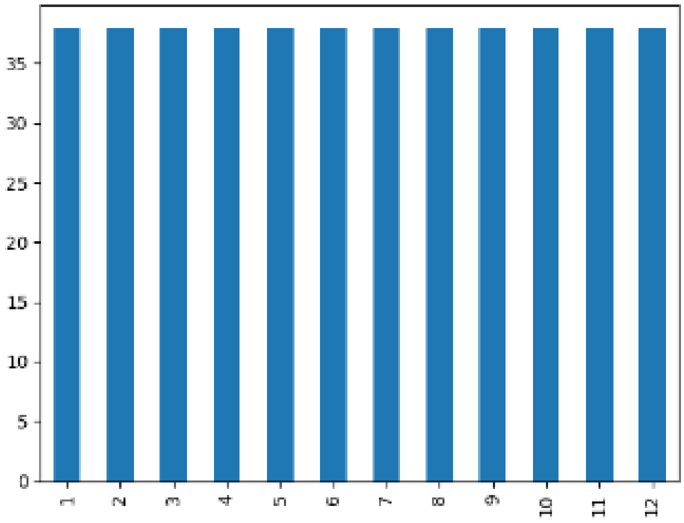


Figure 2 - Plot to display under sample quantities for disorders

The EEG data [13] included 5 minutes of eyes-closed resting state with 19 channels acquired via 500 – 1000Hz sampling rate with 0.1 to 100 on-line filters with Neuroscan [67] [51]. The researchers managed to keep electrode impedances below 5 k by applying electrical conductivity gel. 19 channels (FP1, FP2, F7, F3, Fz, F4, F8, T7, C3, Cz, C4, T8, P7, P3, Pz, P4, P8, O1, and O2) were selected with a basis of the 10 – 20 system [68] in conjunction with a mastoid reference electrode. [69]

FP1 – Marks site for left frontal pole EEG electrode.

FP2 - Marks site for right frontal pole EEG electrode.

F7 – Marks site for left prefrontal, dorsolateral cortex EEG electrode.

F3 – Marks site electrode placed at intersection between Fz-F7 and C3-Fp1 sites.

Fz – Marks site for the midline frontal EEG electrode.

F4 – Marks site for right frontal lobe brain EEG electrode.

F8 – Marks site for right prefrontal, dorsolateral cortex EEG electrode.

T7 – Marks site for the electrode placed halfway between the Fpz and Oz electrodes along the horizontal line over the temporal lobe.

C3 - Marks site for the electrode placed on the postcentral gyrus.

Cz – Marks site for midline central electrode placement.

C4 – Marks site for electrode placed between the forehead and nose.

T8 – Marks site for the electrode placed halfway between the Fpz and Oz electrodes along the horizontal line over the temporal lobe.

P7 – Marks site for electrode placed at left posterior parietal region of the head (roughly behind the left ear).

P3 – Marks site for electrode placed in left parietal region of head (slightly above and behind left ear).

Pz – Marks site for parietal region of head at the midline (top of the head) electrode.

P4 – Marks site for electrode positioned roughly above and behind the right ear in the parietal area of the head.

P8 – Marks site for electrode placed in the right parietal region of head (slightly above and behind the right ear.

O1 – Marks site for rear left occipital region electrode.

O2 – Marks site for rear right occipital region electrode.

To effectively visualise the placement of the electrodes a package called MNE was used to create a 2D plot of an individual's scalp and highlight each site. A 10-20 system [68] was used for the placement of the electrodes and the reason this is important in data retrieval is with pattern recognition of various wave lengths it can help our understanding of the individual's brain being studied. The signals can then be partnered with the reciprocating disorder to determine if there are unique attributes, features or patterns that can help identify the disorder in other individuals and therefore help the overall diagnostic process. [69]

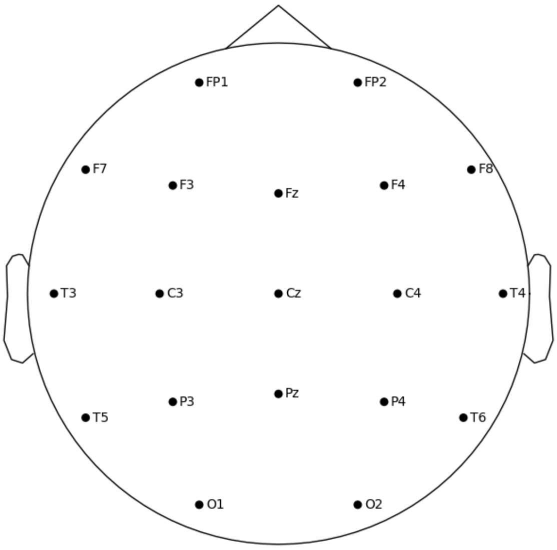


Figure 3 - Visual of electrode placement

Neurosky introduced the concept of EEG [13] frequency bands [67] [51] [69] [69] during 2015, these bands are components of the overall EEG [13] waveform captured at an electrode. FFT models were used to extract the band data from the waveform. When enough neurons oscillate together at a given frequency, the signal can be detected by an EEG. The most common waveforms according to Neurosky are Alpha, Beta, High Beta, Delta, Gamma, and Theta. [67] [51] [69] [69]

Alpha – Signals typically between at 8-12Hz. High volumes of Alpha waves generally indicate daydreaming, an inability to focus and being overly relaxed with low volumes indicating stress, anxiety, insomnia, and OCD. Optimal waves reflect creativity, emotional connection, intuition and possibly relaxation.

Beta / High Beta- Signals typically between at 12-35Hz. High volumes of Beta waves can indicate adrenaline, anxiety, high arousal, or stress with low volumes indicating potential ADHD, possible daydreaming, depression, or even poor cognition. Optimal waves generally reflect conscious focus, memory activation or problem solving.

Delta - Signals typically between at 0.5-4Hz. High volumes of Delta waves can indicate brain injuries, learning problems, inability to think or even severe ADHD. Low volumes of this brain wave can indicate an inability for the body to rejuvenate, inability for the brain to revitalise or poor sleep. Optimal Delta wave patterns can indicate an active immune system, natural healing, and deep sleep.

Gamma - Signals typically found at 35+Hz. High volumes of Gamma waves are associated with anxiety, high arousal with low volumes indicating potentials ADHD, depression, and learning disabilities. Optimal waves typically reflect binding senses, cognition, information processing, general learning, and REM sleep.

Theta - Signals typically between at 4-8Hz. High volumes of Theta waves indicate depression, hyperactivity, impulsivity, inattentiveness, and ADHD. Low volumes reflect anxiety, poor emotional awareness, and stress. Optimal waves reflect creativity, emotional connection, intuition and possibly relaxation.

Using data visualisation tools available to the Python language, and from a source on Kaggle I developed a plot that displays all specific disorders against their waveform activity. This helped to acknowledge how on each waveform there were various levels of activity in relation to the specific disorder recorded in various brain regions. This indicated that with data exploration and the appropriate algorithm that it would be possible to extract a model that offers consistent accuracy when performing classification of the available disorders. To further understand the differentials, I had to investigate the use of several algorithms and architectures against the EEG dataset [13] to extract the maximum reward.

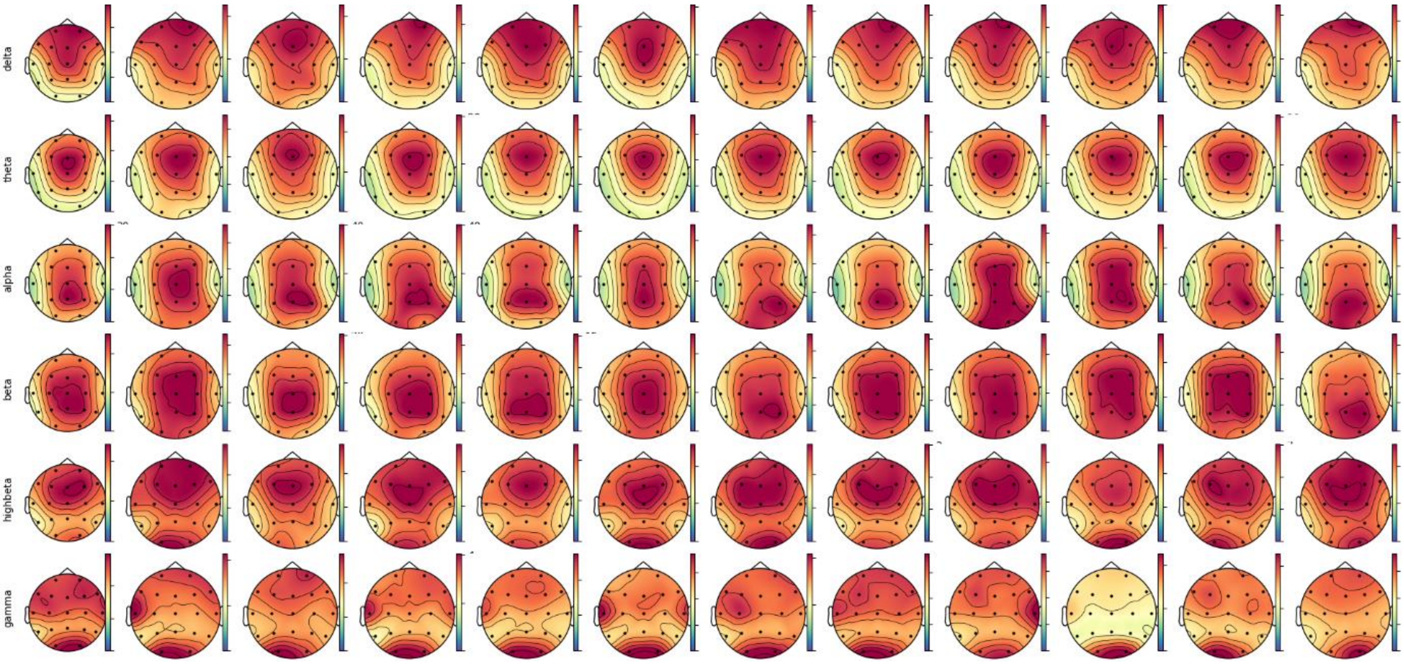


Figure 4 - EEG activity for individual brain waves upon specific disorders

Classification Models

With our model requiring the ability to classify a specific disorder from a range of psychiatric disorders, selecting an appropriate machine learning classification model would be vital. Unlike the published paper in which I obtained the dataset from my research will focus on only using the specific disorder column along with the EEG data [13] as the sole aim of my research is to determine if it is possible to diagnose an individual correctly by only observing their EEG signals [13].

According to Simon Tavasoli [70] (data science advisor and instructor) and Rob Schapire (Princeton University) the best supported ML [71] [22] [72] algorithms for classification are [73] Logistic Regression, [74] Decision Trees, [17]SVM, [75] Random Forests, [38] Naïve Bayes, and [76] KNN algorithm.

These models fall into three types of ML categories, [77] UL, [78] SSL and [79] Reinforcement Learning.

[77] UL is when ML algorithms analyse and cluster unlabelled datasets, the algorithms detect hidden groups and patterns within the data without human input. UL according to IBM is commonly used in clustering, data association and for dimensionality reduction.

Advantages of UL [77] are the models can handle data of higher dimensions than humans can visualise or process. These models also make it possible to find patterns and relationships in data that is unlabelled and often the outcomes can yield relationships that may not be found in labelled datasets.

Disadvantages of UL [77] often include the model's overall performance is not as accurate as a supervised model as the patterns and relationships found may mislead and skew the model with untrue patterns. It can cost more to run this model type as additional fee may occur for employing a human to intervene and understand the patterns developed by the model to determine if it is trustworthy.

SSL [78] according to Scikit-learn is when a specified amount of training data is not labelled. The SSL algorithms make assumptions from observing the distribution of the dataset to determine patterns in the data and by identifying these patterns the model can achieve performance gains.

Advantages of SSL [78] include improved model learning as it utilises the available labels to help with processing the unlabelled data. Neetika Khandelwal states the model's performance and ability to generalise is easier than that of the unsupervised models. Also, this model with limited labels can complete standard machine learning tasks with state-of-the-art results making it a popular choice where datasets contain limited amounts of labels. [77] [79]

Disadvantages of SSL [78] can consist of unreliable results when producing labels for unlabelled data even though it may find consistent patterns or relationships. The performance will never be as accurate as Supervised Learning. [77] [79]

SL [80] according to IBM is defined by making use of labelled datasets to train algorithms to either predict or classify outcomes accurately. Input data is directly passed into the model and the model will adjust its weights accordingly until it has been appropriately fitted via cross validation.

Advantages of SL [80] are its ability to develop models that contain exact concepts of the classes in the training data. In terms of learning ML concepts, the supporting article state is it simpler to understand and therefore easier to learn as the observant can see how the relationships in the data are form and what classes are linked together which can be used to derive meaning. This ML type provides a high level of accuracy and is very helpful in classification problems according to Pythonistaplanet. [77] [79]

Disadvantages of SL [80] include an inability to cluster or classify data by finding its own features, which unsupervised learning can. Another downfall is that in the scenario where input data is not in the form of other available classes, the model may incorrectly label the class. Supervised Learning models are data hungry and if an insufficient amount of training data is used the accuracy of the model suffers. Training models for generalisation requires a lot of computation power and time which leads to higher equipment costs as better processing and memory-based units will be required. [77] [79]

Logistic Regression (supervised learning) [73] [81] [82] [37]

Logistic regression (developed by David Roxbee Cox 1958) models are most used in classification and predictive analytics making them perfect for my multiple classification problem. How this model architecture works is that it estimates the probability of a particular outcome for example, did the patient have a disorder or not. With the outcome being based on probability the dependent variable is restricted between 0 and 1. Logit transformation is applied to the odds to calculate the probability of success divided by the probability of failure. The formula is explained below.

'*Logit(pi)*' represents the log-odds of the probability of a binary event happening. The purpose of this function is to transform the probability of '*pi*' into a continuous value.

'exp' stands for the exponential method which in this formula is raised to the power of the argument inside the parenthesis '(-pi)'.

'*1 + exp(-pi)*' is used to denote the denominator of the fraction which is the total of 1 with the exponential of *'(-pi)'*.

'*1 / (1 + exp(-pi))*' is the whole formula to represent Logistic function with the numerator in this case being set to 1 and the denominator set to +1 plus the exponential of *'(-pi)'*. This ensures the output is a probability value of between 0 and 1.

'*In(pi/(1-pi))*' stands for the logarithm of the odds ratio where '*pi*' is the probability of a binary event occurring (1 stands for success and 0 for failure).

'*Beta0, Beta1…. Betak'* represent the weight/feature/coefficients associated with the variables '*X1,X2…,Xk'* for predictions in the Logistic Regression model. Each 'Beta' stands for the effect of 1 predictor variable on the log-odds of success.

$$\text{Logit}(pi) = 1/(1+ \exp(-pi))$$

$$\ln(pi/(1-pi)) = Beta_0 + Beta_1*X_1 + \ldots + B_k*K_k$$

Equation 1 - Logit formula

This formula states that the '*logit(pi)*' is the dependent variable and x is the independent variable with the beta parameter / coefficient being estimated via maximum likelihood estimation (MLE). The aim of the MLE method is to test various values of beta through multiple iterations to optimise for the best fit. All iterations produce the log likelihood function with logistic regression seeking to optimise this function to find the best parameter. With an optimal coefficient located the conditional probabilities can be calculated for each observation, this is then recorded and added together to extract an overall probability prediction. In the instance of a binary classification problem, probabilities greater than 0.5 will predict 1 and vice versa for 0. There are three main types of logistic regression models. [73] [81] [82] [37]

Binary logistic regression – With this approach there are only two possible outcomes being 0 or 1, e.g., determining if an individual is sick or not which is a binary classification problem.

Multinomial logistic regression – This model type is used when there are three or more possible outcomes, e.g., determining what illness an individual is suffering from by calculating the weights of their symptoms.

Ordinal logistic regression – With this logistic regression model the response variable has three or more possible outcomes with the values having a defined order, this may be a useful model type for determining classification of EEG signals to match the best corresponding disorder.

Advantages of logistic regression according to opengenus.org include being one of the easiest machine learning algorithms making it a great training choice for those studying machine learning. The trained features / weights provide insight to the importance of each feature which means we can understand the data by observing the relationships taking place in the data. Updates to this algorithm can be made using stochastic gradient descent making it easier to implement new data. When dealing with a low dimensional dataset with enough training data this model type will also be less prone to over-fitting. This model is very efficient when the dataset contains features that are linearly separable as it effective designs a boundary line that divides data into groupings, another benefit of this approach is that time is saved in compilation as the algorithm's simplicity allows it to converge faster. It has also been stated that resultant weights discovered after training the model are highly interpretable meaning the models pattern selection can be understood by a human and therefore take more information away from the model. [73] [81] [82] [37]

Disadvantages of logistic regression include the potential for overfitting the training set when working with a high dimensional dataset (overfitting is when the model is too confident in its accuracy of predictions on the training set and when shown new data from the test set accuracy can heavily diminish). To combat the potential overfitting regularisation techniques can be used but this then increases the overall model complexity which can make it difficult to understand when producing results. This model type can't process non-linear problems as the core methodology behind this model type is to make a linear decision. This model can also struggle to handle complex relationships of higher dimensions making it unsuitable for problems where possibly thousands of variables need processed. Another downside to logistic regression is its requirements to have either moderate or multicollinearity between independent variables (this means if a few variables share a high correlation, only one should be selected as repetition of data may lead to a bias model that trained on inappropriate training parameters). If the data being handled contains outliers the algorithm will process them as equals therefore skewing results. [73] [81] [82] [37]

Elastic Net (supervised learning) [83] [84] [85]

Elastic net (developed by Zou and Hastie 2005) is a modified regularised linear regression that makes use of two additional penalties to the loss during training, these additions being L1 and L2 methods. This algorithm addresses the problems of stability in linear regression as it is sensitive to inputs, to counter this the loss function is modified to include additional costs for a model with large coefficients.

The first method for penalizing a model is the use of L1 which is based on the sum of the absolute weight/feature values minimises the size of all coefficients and only allows a few values to be minimized to 0 removing the predictor from the model.

The ‘l1_penalty’ represents L1 regularisation also referred to as Lasso is used to add the regularisation term to Linear Regressions model’s cost function with the aim to prevent overfitting by adding a penalty to absolute values of the models’ weights/features/coefficients.

‘*sum j=0 to p’* represents the totalling up of regularisation terms for each weights/features/coefficients from *‘j=0’* to ‘*j=p’*.

‘*beta_j’* is used to stand for the n-th weight/feature/coefficient of the Linear Regression model as each weight/feature/coefficient has its own beta weight/feature/coefficient.

‘*abs(beta_j)*’ represents the absolute value of each weight/feature/coefficient. It is at this stage absolute values are converted from negative to positive values so that the L1 penalty calculation is based on the magnitude of weight/feature/coefficients.

‘*sum= j=0 to p abs(beta_j)’* is the equation that is the sum of the absolute values of the weight/feature/coefficients for all weight/feature/coefficient (from j=0 to j=p). For regularisation the sum is added to the Linear Regression cost function with the goal of finding the values of weight/feature/coefficients that minimise the sum of squared differences between predicted and actual values of training data.

l1_penalty = sum j=0 to p abs(beta_j)

Equation 2 - L1 penalty equation

The second method for penalizing a model is the use of L2 which is based on the sum of the squared coefficient values. This method not only minimises the size of the coefficients but also prevents any coefficients from being removed from the model. [83] [84] [85]

‘l2_penalty’ represents L2 regularisation and is also referred to as Ridge regularisation and is used to prevent overfitting and stabilize the model by adding a penalty for large values for weight/feature/coefficients.

‘*sum j=0 to p’* represents the totalling up of regularisation terms for each weights/features/coefficients from ‘*j=0’* to ‘*j=p’*.

‘*beta_j’* is used to stand for the n-th weight/feature/coefficient of the Linear Regression model as each weight/feature/coefficient has its own beta weight/feature/coefficient.

‘*beta_j^2’* represents the square of each beta weight/feature/coefficient. By squaring the weight/feature/coefficients the penalty is larger which encourages the model to prefer smaller weight/feature/coefficients during optimisation.

‘*sum j=0 to p beta_j^2’* is the formula to represent the total of squared weights/features/coefficients for all weights/features/coefficients from ‘*j=0 toj=p’*. The aim is to find the values of beta that minimise the total/sum of squared differences between the predicted and actual values whilst also minimising the penalty term.

l2_penalty = sum j=0 to p beta_j^2

Equation 3 - L2 penalty equation

An Alpha value is applied to both Lasso/L1 and Ridge/L2 penalties to determine the weight/influence of the results of each loss function as an example if the Alpha value is set to 0 then all of the weight is assigned to the L2 penalty and vice versa for L1. [83] [84] [85]

Advantages of Elastic Net include the ability to work with multicollinearity which is when weights are highly correlated with one another. The algorithm can reduce overfitting via utilising the benefits of both the Lasso and Ridge methods for adding regularisation to the model by balancing the trade-off between underfitting and overfitting. The algorithm can also perform feature selection which is when the model identifies the most significant weights/features for the overall outcome of the model and it does this by setting some weights/features to 0. [83] [84] [85]

Disadvantages of Elastic Net include the requirement for hyper tuning two vital parameters, Alpha and Lambda. These values need specified by the engineer of the model and on a large scale this can significantly impact the model’s overall performance, the cost to develop and heavily increase the time required for development. Elastic net may not also perform well when handling high dimensional data when the number of features is larger than the number of observations as the algorithm faces difficulty when selecting relevant features meaning it can’t reduce the dimensionality efficiently. Finally, this algorithm can be treated as black box as the model’s selection of features with high dimensionality is difficult to interpret and explain. [83] [84] [85]

Decision Trees (supervised learning) [86] [74]

Decision trees (developed during THAID project by Mandell 1972) are commonly used for either classification or regression tasks. It makes use of a hierarchical tree-based structure which can be denoted as containing a root being the node, branches, internal nodes, and leaf nodes.

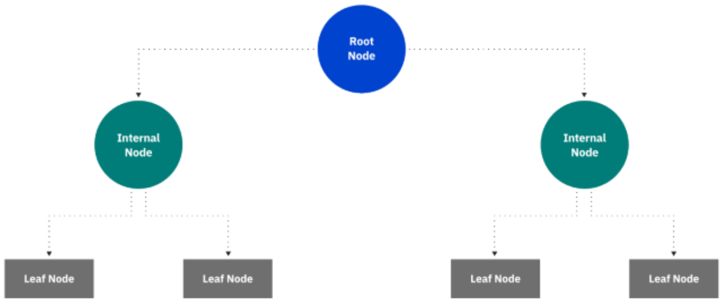


Figure 5 - Decision tree graph

As shown in the diagram above Decision trees begin with a root node and the outgoing branches from the root node feed data into the decision/internal nodes. With reference to the features/weights both node types perform evaluations to form subsets in the data which are referred to as leaf/terminal nodes.

A greedy search method is performed to identify the optimal divide in the data, this process is then reiterated from top to bottom until either all or a large percentage of the data records have been classified. Regarding the model's ability to classify all data majorly depends on the complexity of the decision tree, smaller decision trees can obtain more of the data and produce what are referred to as pure leaf nodes. However, as the tree grows it can become difficult to maintain the purity of the leaf lead and often is results in low amounts of data being passed into the subtree, this fault is known as data fragmentation. Data fragmentation can lead to overfitting which as discussed previously gives the model a false sense of confidence in its ability to process data and when it handles unseen data its accuracy drastically falls. To combat this model's potential overfitting the data pruning method is used which removes branches that split weights of low importance. According to IBM an alternative way to maintain a decision trees accuracy is by forming an ensemble via the random forest algorithm which can predict more accurate results with success when individual trees are uncorrelated. There are three main types of Decision tree. [86] [74]

ID3 – Iterative Dichotomiser 3 developed by Ross Quinlan 1986 utilises entropy and data gain as its main metrics to evaluate splits in the data. This algorithm tends to lead to multiple trees and has difficulty handling missing or noisy data.

C4.5 – This algorithm is believed to be a modern iteration of ID3 which was also developed by Ross Quinlan, it uses data gain or gain ratios to evaluate divides within the decision tree. Unlike ID3 C4.5 can handle missing data and addresses the issues of overfitting in ID3 via pruning (removal of branches).

CART – The algorithm developed by Leo Breiman makes use of Gini impurity (measures how regularly a random attribute is misclassified, lower Gini values are more ideal) to identify the optimal attribute to split the data on.

C4.5, ID3 and CART starts with a dataset containing instances and target variable, next it selects the best attribute to split the data based on information gain which measures the reduction in entropy (uncertainty) for the target variable upon selecting an attribute to split the data.

For the tree construction this algorithm creates a node for the decision tree by utilising the selected attribute as a decision point. The next step is to split the dataset into subsets based on the potential values of the attribute with each subset corresponding to a branch from the current node in the tree. The previous step is repeated until either all instances in the subset belong to the same class or there are no further available attributes or a stopping statement has been met. With the tree being fully developed the leaf nodes represent class labels and when new data is to be classified it works along the tree until meet a leaf node (represents prediction variable).

Pruning which is an optional step can be applied to reduce the algorithms complexity and to combat overfitting as the process involves removing branches that do not impact the model with a great significance.

Advantages to using the decision tree algorithm are it is easy to interpret as it makes use of Boolean logic and has a lot of support packages to visualise a representation of the tree. The hierarchy structure of a decision tree makes it easy to derive information about the tree and which attributes are valued as the most important in dividing the dataset. This model type according to IBM is more flexible when compared to other classifiers as it requires very little data preparation as it can handle continuous numbers and even deal with missing values. The model is also insensitive to underlying relationships between weights meaning if there are two weights that are highly correlated the algorithm is smart to select only one to split the weights on. [86] [74]

Disadvantages of complex decision trees includes a vulnerability to overfitting which means the model cannot generalise effectively with new data, to combat this pre/post-pruning can be used to either stop the tree growth if there is insufficient data whereas post-pruning removes sub trees with insufficient data. This algorithm is also prone to high variance estimations especially when small variations in the data are made, these changes can lead to very different results. In terms of cost-effective algorithms this can suffer from an expensive training phase whilst using the greedy search method. Finally, this algorithm is not fully supported by scikit-learn who are responsible for handling a large machine learning library in python which makes it difficult to fully trust. [86] [74]

XGBoost (supervised learning) [87] [88] [89] [41]

XGBoost (developed by Tianqi Chen and Carlos Guestrin 2016) is an implementation of gradient boosting on decision trees to improve speed, ease of use and performance for large scale datasets. XGBoost makes use of regularisation to control underfitting and overfitting via the L1/L2 penalties on weights/features and can support both regression and classification problems.

XGBoost can be broken down into individual steps, first beginning with initialisation is when the model is set with a constant prediction whilst calculating the initial residual errors. Tree development is next and for each iteration a decision is built that predicts the negative gradient of the loss function, this tree essentially becomes the base learner. The residuals are then updated by subtracting the predictions of the model against the true value. Next L1 / L2 regularisation techniques are applied to determine the complexity of the tree with the aim of combatting overfitting. The trees are continuing to be built until a stopping criterion is met for example it could be a maximum number of trees or a certain improvement rate in the loss function. Finally, the prediction is made for new data points by being evaluated against all the trees in the ensemble and totally up their individual predictions to make the classification. The formula is explained below.

‘*L(t)*’ is the objective function (quantifies the performance of the models’ predictions against the target values) with the iteration of ‘*t*’ from the training process.

‘*l(yi, yˆi (t−1) + ft(xi))*’ represents the loss function for the ‘*i'th* data point at iteration ‘*t*’. This measures the discrepancy between the actual target value *‘yi’* and predicted value ‘"*yˆi (t−1) + ft(xi)*’. ‘*yi*’ is the true target value for data point.

‘*yˆi (t−1)*’ is the predicted value for the ‘*i'th* data point at the previous iteration ‘*t-1’*.

‘*ft(xi)*’ contains the contribution of a new Decision Tree to the prediction for the ‘*i'th* data point on its features.

‘Ω*(ft)*’ this term represents the regularisation applied to the learner ‘*ft*’. This is used to control the complexity of the model and prevent overfitting.

$$\mathcal{L}^{(t)} = \sum_{i=1} l\left(y_i, \hat{y_i}^{(t-1)} + f_t(\mathbf{x}_i)\right) + \Omega(f_t)$$

Can be seen as f(x + Δx) where x = $\hat{y_i}^{(t-1)}$

Equation 4 - XG Boost regularisation with l1 and l2 penalties for weights

Advantages of XGBoost include a reduction of the feature engineering as there is no requirement for scaling, normalising data or handling of missing values which can be difficult and time consuming. Another benefit of this algorithm is it allows access to feature/weight importance which is useful for understanding the model and where it has based most of its prediction work on. The algorithm can handle outliers and converge in a timely manner with large datasets. [87] [88] [89] [41]

Disadvantages of XGBoost include the process required for visualisations of the model is difficult to achieve. Overfitting is possible if the L1/L2 penalties are incorrectly tuned which means model performance can suffer. [87] [88] [89] [41]

<u>LightGBM (supervised learning)</u> [90] [36] [91] [92]

LightGBM (developed by Mercedes Erra and Jurgen Schmidhuber 2016) according to Pushkar Mandot is a gradient boosting framework that utilises the decision tree algorithm. The decision tree algorithm is modified to develop trees in a vertical manner versus the usual horizontal one which means the tree grows leaf wise instead of level wise. LightGBM selects the leaf with the maximum delta loss in order to grow.

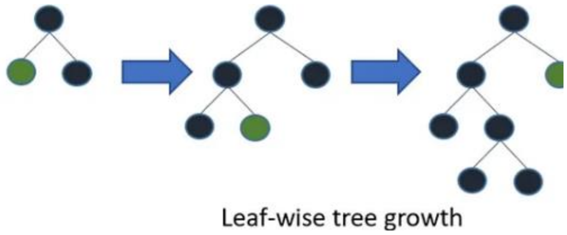


Figure 6 - Example of how LightGBM using Decision tree that work vertically

Below the formula is explained for LightGBM [90] [36] [91] [92]. ‘n’ is the total quantity of data points in the training dataset. ‘*i*’ is the variable to symbolise the individual data points in the training data. ‘*yi*’ is the true value for each of the data points. ‘*y^i*’ is the predicted value for each data point. ‘*k*’ is the number of leaves /nodes within the ensemble of trees. ‘*fk*’ is the score associated with the ‘*k-th’* leaf to indicate the level of contribution or weighting of data assigned to that leaf. ‘*L(yi,y^i)*’ represents the loss function to measure the error rate between the true values and predicted values. ‘*Ω(fk)*’ is the regularisation term to penalise complex trees which combats overfitting in the model.

$$\text{Objective} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Equation 5 - Objective formula for LightGBM architecture

When dealing with multiclass classification ‘softmax’ is used to assign probabilities of each class to the EEG data. ‘*k*’ is the number of unique labels for classes. ‘*xi*’ represents the feature/weight/coefficient vector and ‘*yi*’ is the true label for the classes. ’*P*(*yi*=*j*|*xi*)’ is the part of the formula that handles the probability of data point ‘*xi’* which belongs to the class in ‘*j*’. ‘*efj*(*xi*

)’ is used to express the exponentiation of the score ‘*fj*(*xi*)’ which is associated with the class ‘*j*’ for the data point assigned to ‘xi’. *‘∑k=1Kefk(xi)’* expresses the total of exponentiated scores for all classes in variable ‘*k*’ with all the data points of ‘*xi*’. The main purpose of this part of the formula is to ensure that the probabilities sum up to 1 via normalisation (this is the process of transforming the data into a range of between 0 and 1).

$$P(y_i = j|x_i) = \frac{e^{f_j(x_i)}}{\sum_{k=1}^{K} e^{f_k(x_i)}}$$

Equation 6 - Softmax formula for LightGBM

Advantages of LightGBM algorithm include a faster training speed with performance efficiency as the architecture discretises buckets with continuous feature/weight values into bins which help speed up the training process. Another benefit of continuous values being discretised is the lower memory usage which can help costs in terms of running the algorithm in a large scale. The algorithm has proven track record with handling big data according to Subham Surana. [90] [36] [91] [92]

Disadvantages of LightGBM include increased algorithm complexity from dividing the tree leaf’s which can lead to overfitting the data. Another downfall of this algorithm’s potential to overfit can come from handling small quantities of data. [90] [36] [91] [92]

Cat Boost (supervised learning) [92] [34] [93] [94]

Cat Boost is a categorical boosting algorithm that uses balanced decision trees that are symmetric in structure, meaning that in each step the same feature/weight pair that result in minimal loss are chosen to be applied to all nodes of that level. The results of the loss are used in the gradient to update predictions by adding the scaled version of the gradient.

Cat Boost also makes used of the ordered boosting algorithm which optimises the learning objective method by manipulation the features in a specified order that can result in a faster convergence with better modal accuracy.

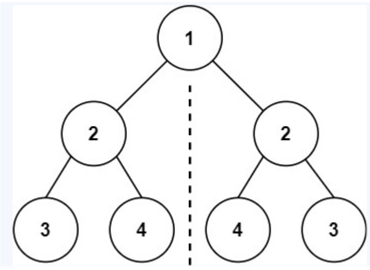


Figure 7- Example of how Cat Boost data is divided

When dealing with multiclass classification ‘softmax’ is used to assign probabilities of each class to the EEG data. ‘*k*’ is the number of unique labels for classes. ‘*xi*’ represents the feature/weight/coefficient vector and ‘*yi*’ is the true label for the classes. ’*P*(*yi*=*j*|*xi*)’ is the part of the formula that handles the probability of data point ‘*xi*’ which belongs to the class in ‘*j*’. ‘*efj*(*xi*)’ is used to express the exponentiation of the score ‘*fj*(*xi*)’ which is associated with the class ‘*j*’ for the data point assigned to ‘*xi*’. *‘∑k=1Kefk(xi)’* expresses the total of exponentiated scores for all classes in variable *‘k’* with all the data points of ‘*xi*’. The main purpose of this part of the formula is to ensure that the probabilities sum up to 1 via normalisation.

$$P(y_i = j|x_i) = \frac{e^{f_j(x_i)}}{\sum_{k=1}^{K} e^{f_k(x_i)}}$$

Equation 7 - Softmax formula for Cat Boost

Advantages of Cat Boost (developed by Yandex and Stanislav Kirillov 2019) include an exceptional ability to handle categorical data like the disorder dataset I am focusing on. Cat Boost has an overfitting detector that stops the training when overfitting is observed and this improves the model’s generalisation and performance making it more adaptable to new unseen data. Cat Boost can also handle big data as it has support for distributed training on multiple computers. [92] [34] [93] [94]

Disadvantages of Cat Boost include the requirement and engineering knowledge to hyper-tune parameters for the algorithm. This algorithm is mainly aimed towards categorical data meaning if the problem does not include categorical data other algorithms may be more suited. [92] [34] [93] [94]

Support Vector Machines (supervised learning) [95] [17]

The SVM (developed by Corinna Cortes and Vladimir Vapnik 1995) algorithm locates a hyperplane in a multi-dimensional space that is used to classify EEG data. The algorithm in particular focuses on a hyperplane that can maximise the distance between data points of different classes. Data points that fall on either side of a hyperplane can be assigned to different classes and number of dimensions of a hyperplane depend on the number of features/weights/coefficients.

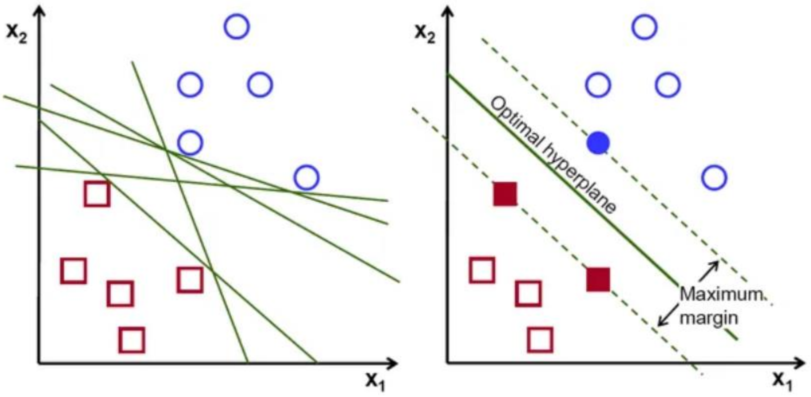


Figure 8 - SVM Hyperplane example

Unlike logistic regression where the output of the linear function is within the range of 0 – 1 using the sigmoid function, SVM takes the output of the linear function and if the output is greater than 1 it is identified with one class and if it is -1 it is identified with another class meaning in SVM the range of values is -1 - 1. With the aim of maximising the margin between data points and the hyperplane a loss function referred to a hinge is used (below is the formula and steps of SVM binary classification explained). [95] [17]

‘*c(x,y,f(x))*’ represents the loss/cost function that measures the discrepancy between the predicted value and the true value.

‘*1-y*f(x)*’ calculates the raw loss value which determines the difference between 1 and the resulting product of true label ‘*y*’ with the predicted value at ‘*f(x)*’.

*‘(1-y*f(x))+’* represents the max function that ensures the loss function does not result in a non-negative. It ensures the value is positive or set to 0, this type of loss is referred to as Hinge.

$$c(x,y,f(x)) = (1 - y * f(x))_+$$

Equation 8 - Hyperplane loss function

If the predicted value and actual value of the same sign, then the cost will be 0 and if not, then a calculation is performed to determine the loss value. A regularisation parameter called the cost function is added to balance the margin maximisation with the loss. [95] [17]

‘min_w’ represents minification of the expression with respect to the weights/features/coefficients vector ‘*w*’.

‘$\lambda \ ||w||^2$’ is the L2 regularisation method of the objective function. Lambda is the hyperparameter which controls the strength of the regularisation and encourages the weights/features/coefficients to be kept small to combat overfitting.

‘$n\sum_{i}=_{1}$ *(1 -* $y_i(x_i,$ *w))_+*’ is the representation for the Hinge loss for data points in the training dataset. ‘n’ is the total number of data points, ‘$\sum_{i}=_{1}$’ is the totally over all data points with the range set by ‘*I*’, ‘*1 -* $y_i(x_i,$ *w)*’ is the calculation of the raw loss for the ‘*i-th*’ data point. ‘*yi*’ is the true label for the ‘*i-th*’ data point and ‘*xi*’ is the feature vector for that particular data point which is then used to make a prediction with the weight vector ‘*w*’. The expression ‘*1 -* $y_i(x_i,$ *w)*’ measures how far the prediction is from the correct side of the decision boundary, if the prediction is on the correct side this term is set to 0.

'*(1 -* $y_i(x_i,$ *w))_+*’ part of the formula represents the Hinge loss for the ‘*i-th*’ data point, it is the max function applied to ‘*1−yi(xi,w)*’. If the prediction is correct then the Hinge loss is 0 otherwise the loss term is positive and contributes to the overall cost function.

$$min_w \lambda \parallel w \parallel^2 + \sum_{i=1}^{n}(1 - y_i\langle x_i, w\rangle)_+$$

Equation 9 - Formula to calculate loss function

With the loss function a subset of derivatives with respect to the features/weights are extracted to find the gradients, with use of the gradients the weights can be updated. [95] [17]

‘$\lambda \ ||w||^2$’ is the L2 regularisation method of the objective function. Lambda is the hyperparameter which controls the strength of the regularisation and encourages the weights/features/coefficients to be kept small to combat overfitting.

$$\frac{\delta}{\delta w_k}\lambda \parallel w \parallel^2 \quad = 2\lambda w_k$$

$$\frac{\delta}{\delta w_k}\left(1 - y_i\langle x_i, w\rangle\right)_+ \quad = \begin{cases} 0, & \text{if } y_i\langle x_i, w\rangle \geq 1 \\ -y_i x_{ik}, & \text{else} \end{cases}$$

Equation 10 - Formula to calculate loss function using gradient weight updates

Once the model can correctly predict the class of the data point, the model will require only updating the gradient from the regularisation parameter. [95] [17]

‘*w*’ is the current weights/features/coefficient vector that is actively being updated. ‘*a*’ is the learning rate (it is typically set to a small integer value to ensure optimisation converges to a minimum without over shooting). ‘*2λw*’ stands for the gradient of the L2 regularisation term with respect to ‘*w*’.

$$w = w - \alpha \cdot (2\lambda w)$$

Equation 11 - Formula to update gradient with regularisation parameter

If there is a misclassification with the prediction of a class, the loss and regularisation parameter is included in the algorithm to perform the gradient update. [95] [17]

‘*w*’ is the current weights/features/coefficient vector that is actively being updated. ‘*a*’ is the learning rate (it is typically set to a small integer value to ensure optimisation converges to a minimum without over shooting). ‘*yi*’ is the target/actual value for the ‘*i-th*’ data point, ‘*xi*’ is the weights/features/coefficients input vector for the ‘*i-th*’ data point. ‘*2λw*’ stands for the gradient of the L2 regularisation term with respect to ‘*w*’.

$$w = w + \alpha \cdot (y_i \cdot x_i - 2\lambda w)$$

Equation 12 - Gradient formula update calculation

With the problem of my project being a multiclass classification issue the SVM was extended with the OVR strategy. In OVR the binary classifiers are trained and each binary classifier is responsible for determining the single disorder against the rest.

$$f_k(x) = w_k^T \cdot x + b_k$$

Equation 13 - OVR decision method formula

‘*fk*(*x*)’ represents the decision function for the number of binary classifiers. EEG data points are assigned to a class ‘*k*’ when processed through the sign ‘*fk*(*x*)’ and if the EEG data point was greater than 0 then it was assigned that specific class otherwise its assigned not that class. ‘*wk’* denotes the weight/feature/coefficient vector that is specific to the ‘*k-th’* binary classifier and it assigns the orientation of the decision boundary for the disorder class. ‘*x*’ is the symbol used to represent the weight/feature/coefficient vector of the EEG data point being classed. Lastly, ‘*bk’* represents the intercept/bias term for the ‘*k-th’* binary classifier and its purpose is to allow for the decision boundary to manoeuvre for the class set in ‘*k*’.

Advantages of SVM according to Scikit-learn are the model is effective in high dimension spaces, even when the number of dimensions is greater than the number of samples. The algorithm is memory efficient as it makes use of support vectors which are developed by using training data in the decision function. The model is versatile as there are a range of Kernel functions that can be utilised for the decision function which could lead to better generalisation. [95] [17]

Disadvantages of SVM show when the number of features/weights much exceeds the number of samples additional work is required to avoid over-fitting such as implementing a kernel function and regularisation term. SVM’s do not have the ability to provide probabilities by themselves and require the use of five-fold cross-validation to produce which is an intensive process, making the model not always computationally affordable. [95] [17]

<u>Random Forests (supervised learning) </u>[75] [96] [39]

Random Forest (developed by Leo Breiman and Adele Cutler 2001) is a method to combine outputs of multiple decision trees to achieve a final single result. This algorithm can handle both regression and classification problems making it extremely resourceful for my EEG dataset. The algorithm extends the bagging method as it makes use of feature/weight randomness to develop an uncorrelated forest of decision trees. This is the key difference between random forest and decision trees, the random forest will make use of only a subset of features/weights whereas decision trees consider all possible feature/weight splits when processing the data.

Multiple decision trees form what’s referred to as an ensemble in the algorithm. Ensembles are used to predict more accurate results when individual trees are not correlated with one another. [75] [96] [39]

The three main hyperparameters of Random Forest algorithms are the node size, number of trees and the number of features/weights to be sampled. To process data the algorithm will first set aside a specified amount of data as test data, next an instance of randomness is injected via feature/weight bagging which adds more diversity in the dataset to reduce correlation amongst decision trees. Next, depending on the type of problem the algorithm will be leveraged in various ways, for classification a most frequent categorical variable is used to retrieve the predicted class and for regression the individual decision tress are averaged to obtain the predicted class. To finalise the out-of-bag sample is used for cross-validation to finalise the prediction.

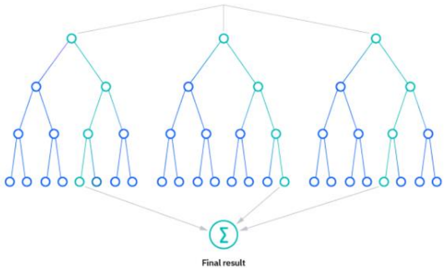


Figure 9 - Random Forest visual graph representation

As the project requires a classification algorithm to be applied the majority vote strategy is used to determine final class labels for EEG data points. The class with the most votes amongst the trees is considered final prediction and can be represented by the formula below.

$$C_i = \operatorname{argmax}_k \left(\sum_{j=1}^{N} \delta(C_{ij} = k) \right)$$

Equation 14 - Random Forest classification formula

'*k*' represents the classes along with 'n' for the trees of the Random Forest. '*Cij*' symbolises the prediction for the class 'i-th' EEG data point by the '*j-th'* tree. The 'Ci' is the final prediction for the EEG data point provided with the small '*k*' representing the individual class. '*δ*(*Cij*=*k*)' is the function of Random Forest that sets the value to either 1 if '*Cij*=*k'* or 0 if it doesn't. The final outcome class is determined by counting how many trees can predict each class (the class/label/disorder) with the most votes wins.

Advantages to the Random Forest algorithm include the reduced risk of overfitting as the classifier won't have the ability to overfit as the model averages the uncorrelated trees which lowers overall variability and prediction error. The algorithm offers the ability to be flexible as it can handle both regression and classification tasks. Feature/weight bagging is another benefit of this algorithm's infrastructure as it can estimate missing values which can therefore retain accuracy in the data. Just like decision trees retrieving feature importance is easy as the Gini importance or decrease in impurity methods can be used to measure the models decrease in accuracy as set variables are excluded. [75] [96] [39]

Disadvantages to the Random Forest algorithm include the slow computation times with large datasets as they compute for each individual decision tree. Another issue of the model is when handling larger datasets additional storage resources will be required to handle the data. When compared to low quantity decision trees it can be difficult to interpret with the increased divisions and complexities. [75] [96] [39]

Naïve Bayes (supervised learning) [97] [97] [98] [99]

Thomas Bayes developed the Naïve Bayes formula between 1702-1761 as an algorithm to deal with classification tasks. This algorithm according to IBM is based on the Bayes Theorem allow for the ability to invert conditional probabilities (conditional probabilities represent the probability of a particular event given another event has occurred). The formula below represents Bayes Rule theorem.

$$P(C|X) = \frac{P(C) \cdot P(X|C)}{P(X)}$$

Equation 15 - Naive Bayes Formula

'*P(C|X)*' represents the posterior probability of the disorder/class with dependence on the features/weights/coefficients '*X*' which are the EEG data values and coherence values in this project. '*P*(*C*)' represents the prior probability of the disorder/class (essentially making an early assumption before processing the data). '*P*(*X*|*C*)' is for the probability of observing '*X*' with relevance to the class/disorder '*C*', it is during this stage where Naïve Bayes makes its assumptions that the features/weights/coefficients are independent given the disorder/class (this is calculated as the product of individual feature probabilities). '*P*(*X*)' represents the probability of observing the features/weights/coefficients '*X*' in the data to perform normalisation which ensures the posterior probabilities total up to 1 with all possible classes/disorders. *'(P(C)·P(X|C))'* represents the predicted class/disorder which gets divided by the probability of the observational features/weights/coefficients.

Bayes theorem is distinct from others as it makes use of sequential events allowing data later implemented to impact the probability of a particular outcome. The prior probability represents the initial probability of an event before conditions are applied sometimes also called the marginal probability and posterior probability is the probability post viewing/observing additional data.

When utilising the Naïve Bayes theorem, it is assumed that the features/weights are either unrelated or conditionally independent and it assumes that all features contribute equally to an outcome. IBM have noted three main types of Naïve Bayes classifiers. [97] [97] [98] [99]

Gaussian Naïve Bayes – This variant focuses on using Gaussian distributions with continuous variables. The models are fitted by extracting the mean and standard deviation of each class.

Multinomial Naïve Bayes – This classifier deals with multinomial distributions which is useful when handling discrete data like integers and numbers.

Bernoulli Naïve Bayes – This version of Naïve Bayes was designed to handle Boolean variables which are variables with only two outcomes e.g., true, or false.

Advantages to the Naïve Bayes classifier according to the IBM article are the algorithm is less complex as the parameters are easier to estimate meaning low level data scientists can create understanding from the algorithm. When this algorithm is compared against logistic regression, Naïve Bayes outclasses in terms of speed and efficiency whilst maintaining a respectable level of accuracy when the independent conditions hold during processing the data. Another great benefit of this algorithm is its ability to hand high dimensional space data which for other classifiers is difficult to process. [97] [97] [98] [99]

Disadvantages with Naïve Bayes include a vulnerability to zero frequency which according to IBM is when a categorical variable does not exist within the training set, to combat this Laplace smoothing can be utilised to handle non-existent training data. This algorithm suffers from false assumption of the independents conditions which does not always hold true. [97] [97] [98] [99]

KNN algorithm (supervised learning) [76] [35]

The KNN algorithm developed by Evelyn Fix and Joseph Hodges in 1951 is a supervised learning classifier that makes use of proximity when performing predictions/classifications. When KNN performs a classification problem class labels are assigned based on majority rule which means the label that is most dominantly represented for a given data point is attributed.

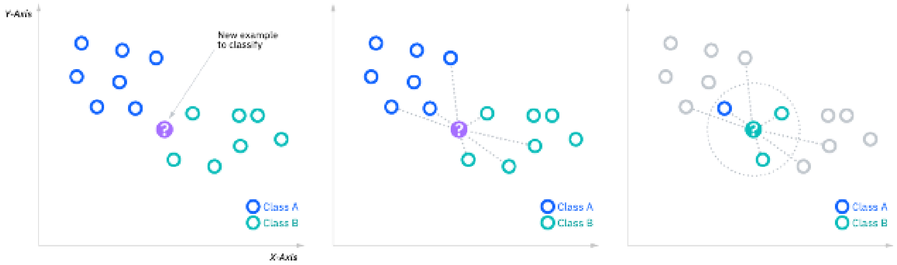


Figure 10 - Example of how K nearest neighbour assigns data

When KKN performs a regression problem the average of the KNN is used to make a prediction on the classification which is a slight variation. The key distinction for this approach is continuous values are used whereas previously discrete values would have been used. Selecting an appropriate algorithm to calculate the distance between a data point and class is vital to the success of this model, in saying that there are multiple options include: [76] [35]

Euclidean distance – This method measures the straight line between the query point versus the identifier for class. ‘*yi-xi’* represents the two points in the space to which the distance between is calculated. ‘*n*’ is the number of dimensions of the feature/weight space. *‘∑i=1’* totals up the results from the calculation for each dimension. *‘(yi-xi)2’* is for each dimension ‘*i*’ to calculate the difference between coordinates of point ‘*y*’ and point ‘*x*’, this distance is then squared to measure the distance along the set number of dimensions. ‘√’ symbol denotes the square root of the total squared differences from the number of dimensions set.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

Equation 16 - Euclidean Distance Formula

Manhattan distance – This method measures the absolute value between two points. ‘*yi-xi’* represents the two points in the space to which the distance between is calculated. ‘*n*’ is the number of dimensions of the feature/weight space. *‘∑=1∑i=1n*’ sums of the absolute differences of the coordinates for each dimension. *‘∣xi−yi∣*’ is for each dimension of ‘*i*’ and the absolute difference between coordinates is calculated.

$$\text{Manhattan Distance} = d(x,y) = \left(\sum_{i=1}^{m}|x_i - y_i|\right)$$

Equation 17 - Manhattan distance formula

Minkowski distance – This method measures the distance between two points by utilising both Euclidean and Manhattan distances as a combination. ‘*yi-xi’* represents the two points in the space to which the distance between is calculated. ‘*n*’ is the number of dimensions of the feature/weight space. *‘∑i=1’* totals up the results from the calculation for each dimension. ‘*p*’ is the positive constant greater than or equal to 1 (referred to as the ‘order’).

$$\text{Minkowski Distance} = \left(\sum_{i=1}^{n}|x_i - y_i|\right)^{1/p}$$

Equation 18 - Minkowski distance formula

For the EEG classification problem, the KNN formula is expressed below. *‘^y*’ symbolises the predicted disorder/class label for the fresh data point set in ‘*x*’. ‘*c*’ is the individual class/disorder label with ‘*K*’ representing the nearest points or neighbours to consider when classifying. ‘*yi*’ represents the disorder/label of the ‘*i-th’* nearest datapoint/neighbour. ‘*δ(yi=c)’* is the function that equals 1 if '*yi-c'* and 0 if it does not. 'argmax' determines the class/label/disorder with the highest quantity amongst the '*K*' neighbours.

$$\hat{y} = \operatorname{argmax}_c \left(\sum_{i=1}^{K} \delta(y_i = c) \right)$$

Equation 19 - KNN classification formula

According to IBM KNN has been referred to as a 'lazy learning' model as it stores a training dataset versus undertaking a training stage, this alters how the model operates as all the computation is performed at the classification/prediction stage. The model is also referred to as memory-based learning as it relies on memory to store its training data.

Advantages of KNN according to IBM are the model can adapt easily as new training data can be adjusted for and stored in the training data memory. There are very few hyperparameters as the method only requires a K value along with a distance metric making it a great onboarding algorithm for new machine learning engineers or enthusiasts to learn. [76] [35]

Disadvantages of KNN include being prone to underfitting and overfitting as lower values of K can overfit data with increased dimensionality and higher values of K can underfit the data. KNN does not scale well as it relies heavily on memory resources which can affect both times dedicated to handling the storage and the finances required to run the model as further storage equipment may be required. [76] [35]

Model results and configurations

With the projects short time frame for completion measures were taken to ensure all classifier variations were performed at least once so that there would be an adequate set of results to get insights into what algorithms performed well with the classification task versus those that didn't.

To save time all classifiers were initially performed with the sklearn train_test_split which divides the data into feature training data (EEG data), feature testing data (EEG data), class/label training data (disorder data) and class/label testing data. It was decided initially to split the data into 70% training and 30% test data as an initial approach to determine if classification could be achieved with any of the previously mentioned algorithm. A benefit of splitting the training data into 70/30% included improved times for the classifier to train as there was less data to train from so this reduced the time required to achieve convergence. The 70/30% and 80/20% splits for the data were most common amongst data scientist, this is according to a forum on research gate where a number of comments were made with various opinions on the subject. [97]

When processing the base line architectures, it was determined this was insufficient to find relationships and differentials as the best result achieved an estimated 22.18% (XG Boost) [40] accuracy which is not credible. Under sampling and over sampling techniques were applied to the range of algorithms and finally there was a drastic improvement in the results as 6 out of 9 algorithms achieving above 75% accuracy with the over sampling technique and a peak of 83.4% achieved by SVM.

The next step to improve model performance was to tune the hyper parameters. Hyperparameter tuning requires a lot of prior knowledge and if I the sole engineer of the project were to tune the parameters manually whilst recording the results to determine the best set it would not be possible to meet the project deadline. As a method to reduce time with tuning the parameters manually the Optuna API [95] was used. The Optuna API allowed for a number of trials to be run with a looping mechanism that would adjust the predetermined hyperparameter ranges depending on the results from the previous trials. Ideally for each model the number of trials run should have correlated to the number of parameters and their range of potential options which in some cases is well into the hundreds, but unfortunately a decision was made to save time and limit the trials to 30 per model which did impact the model's performance with 5 algorithms improving accuracy with a peak score of 84.5% by SVM. Cat Boost, Decision Tree, KNN, SVM and XG Boost all had improvements. KNN had a drastic accuracy improvement from an estimated 54.81% accuracy to 82.84% indicating that hyperparameter tuning is worth exploiting with further trials as part of the project's future works. In some cases, with hyperparameter tuning there were decreases in performance such as Random Forest plummeted from 79% to 24.8% and it was speculated that this could be a result of poor hyperparameters being suggested by the limited number of trials.

As the development of the classifiers came at a later stage in the project it was noted that they were all developed with the train_test_split method [97] with a random_state parameter set with a value of 42, this meant every time the classifier was trained it was trained from new but with the same 70% of training data to allow for consistency which made it possible to assess results at all stages of the models development from base architectures to architectures where over sampling and hyperparameters were tuned without bias in the training data. One downfall of this is the classifiers true potential may not yet have been realised as the 30% of data used for testing may have contained further information which could help in the classification of psychiatric disorders.

Table 1 - 70/30 Training Test split classifier results

Model Architecture	Base Model Accuracy	Under Sampling Accuracy	Over Sampling Accuracy	Over Sampling Hyper Tuned Accuracy
Cat Boost	21.478%	16.788%	**81.589%**	**82.984%**
Decision Tree	12.323%	11.678%	50.069%	51.324%
K Nearest Neighbour	16.901%	10.218%	54.811%	**82.845%**
Light GBM	20.070%	13.138%	**81.450%**	79.218%

Logistic Regression	19.718%	17.518%	80.613%	79.916%
Naïve Bayes	8.450%	15.328%	25.104%	21.338%
Random Forest	19.014%	6.569%	79.079%	24.825%
Support Vector Machine	17.605%	09.489%	**83.403%**	**84.518%**
XG Boost	22.183%	12.408%	77.405%	78.940%

The solution to this was to implement repeated K fold cross validation [97] [98]. Repeated K-fold cross validation is a strategy to improve the classifiers estimated performance as it repeats the cross-validation method several times to compute a mean result across all K folds. Benefits of repeating the K-fold cross validations include a method to reduce the error rate when estimating the model's performance and it negates the potential for noise during the training data split.

Being conscious of time the number of folds for cross validation was limited to 5 giving a training and test split of 80/20% which from opinions on ResearchGate seems to the most common mentioned with repeats set to 5, this means the cross validation with the each algorithm trial was repeated 5 times (ideally this would have been increased to 10, but computation time for 9 architecture types required too much time), this helped to reveal the mean for model accuracy more precisely and the standard error that is estimated to be expected from the model. 6 out of the 9 models achieved a mean accuracy score of 80+% with additional gains in terms of classifier performance as the training data size was increased allowing the classifiers to find further discriminations between disorders. Cross-validation was successful in its objective which was to increase performance gains whilst repeating the folds of the training data set to ensure the original train test split did not contain bias. KNN with the 'Minkowski' setting achieved the second highest accuracy score (84.82%) indicating a much simpler classifier can perform just as well as the more sophisticated and supported architectures. A simpler classifier also means computation time to achieve convergence is reduced making it a valid classifier to investigate further for generalisation as a faster model means overall resources saved. It had been noted there were few articles which had explored KNN as a means to classify psychiatric disorders via EEG making this research more valuable.

standard_error = sample_standard_deviation / sqrt(number of repeats)

Equation 20 - Standard error

The standard error equation Equation 20 provides an indication as to the amount of error that can be expected from the training sample mean. The number of repeats selected should reflect in the minimisation of the standard error and a lower standard error score reflects in model performance stability.

Table 2 - 5 K fold cross validation mean results and mean standard error rate

Model Architecture	5 K Fold Mean Accuracy Score	Mean Standard Error Rate
Cat Boost Hyper tuned with Optuna	**83.443333%**	0.04%
Decision Tree Hyper tuned with Optuna	53.498%	0.07%
K Nearest Neighbour Hyper tuned with Optuna	**84.822%**	0.044%
Light GBM Hyper tuned with Optuna	**83.17%**	0.044%
Logistic Regression Hyper tuned with Optuna	**82.3525%**	0.06%
Naïve Bayes Hyper tuned with Optuna	23.62%	0.048%
Random Forest Hyper tuned with Optuna	26.89%	0.078%
Support Vector Machine Hyper tuned with Optuna	**86.474%**	0.04%
XG Boost Hyper tuned with Optuna	**80.788%**	0.058%

As part of data exploration and understanding the SHAP API [100] was used to perform permutations and extract the most meaningful / heaviest weighted features in the models. With the SHAP API [100] identifying the most important features for each classifier and they were arranged them from most important to least, the reason for this was to perform an experiment with the top 10 features for each classifier as shown in Table 3 below. Each model would be rerun with only these features to see if an overall improvement on accuracy whilst reducing the time for convergence with significantly less data was possible.

Unfortunately, this experiment was not successful when handling the 10 most important features as the best performing algorithm was Logistic Regression achieving 22.18% accuracy reflecting in a failed experiment. As a result, the number of features were increased to 100 as a strategy to yield more meaningful features.

Table 3 - Accuracy results from utilising 10 best features extracted by SHAP API

Model Architecture	Base Model Accuracy	Under Sampling Accuracy	Over Sampling Accuracy	Over Sampling Hyper Tuned Accuracy
Cat Boost	13.380%	17.253%	15.492%	20.774%
Decision Tree	12.323%	12.676%	16.901%	14.788%
K Nearest Neighbour	12.676%	12.676%	15.492%	18.309%
Light GBM	15.845%	14.436%	17.253%	18.309%
Logistic Regression	**21.126%**	**21.126%**	**21.126%**	**21.478%**
Naïve Bayes	10.915%	12.323%	10.915%	10.915%
Random Forest	**23.239%**	**21.478%**	18.661%	**22.183%**

Support Vector Machine	20.422%	21.478%	Requires further computation time	Requires further computation time
XG Boost	15.845%	17.253%	17.605%	19.366%

This experiment also failed with the best performing classifier Cat Boost only achieving 22.535% accuracy as highlighted in the comparison Table 4 below. With more time spent on the investigation of the features selected by SHAP and further increasing the number of features for each classifier until achieving the same levels of accuracy in the previous table and to understand what features are redundant to reduce overall computation expenditure whilst improving convergence times requires further trial runs. As a result of the best features performing poorly when the classifier was retrained on only the selected features it was determined that passing comment or observations was the best course until further experiments were carried out as part of future project work.

Table 4 - Accuracy results of models hyper tuned with 10 vs 100 best features

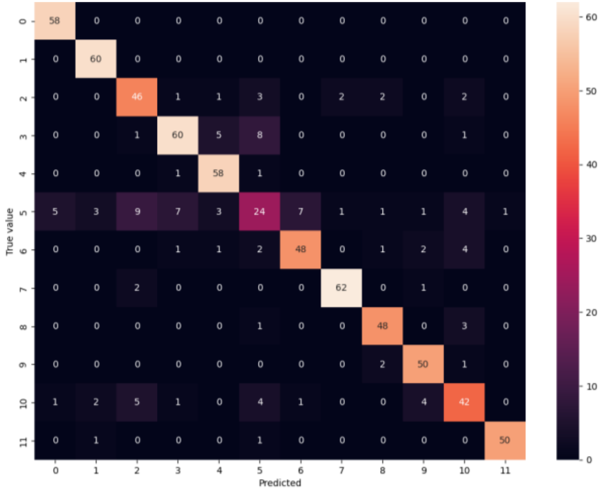
Model Architecture	Over Sampling Hyper Tuned Accuracy 10 Features	Over Sampling Hyper Tuned Accuracy 100 Features
Cat Boost	20.774%	**22.535%**
Decision Tree	14.788%	13.380%
K Nearest Neighbour	18.309%	10.915%
Light GBM	18.309%	**22.183%**
Logistic Regression	**21.478%**	19.014%
Naïve Bayes	10.915%	10.211%
Random Forest	**22.183%**	**21.478%**
Support Vector Machine	Requires further computation time	Requires further computation time
XG Boost	15.845%	21.126%

Confusion matrix feedback analysis

The confusion matrix [101] was used to derive additional understanding from the classifiers strengths and weaknesses when assign each disorder/class label. The confusion matrix was created by applying the true (rows) value against the predicted value (columns) of the model to evaluate areas of confusion. All values off of the diagonal represent an error / confusion by the model. To help understand what numbers reciprocate to what disorder the following list should help:

- 0 = Acute Stress disorder
- 1 = Adjustment disorder
- 2 = Alcohol use disorder
- 3 = Behavioural addiction disorder
- 4 = Bipolar disorder
- 5 = Depressive disorder
- 6 = Healthy control
- 7 = Obsessive compulsive disorder
- 8 = Panic disorder
- 9 = Post traumatic stress disorder
- 10 = Schizophrenia disorder
- 11 = Social anxiety disorder

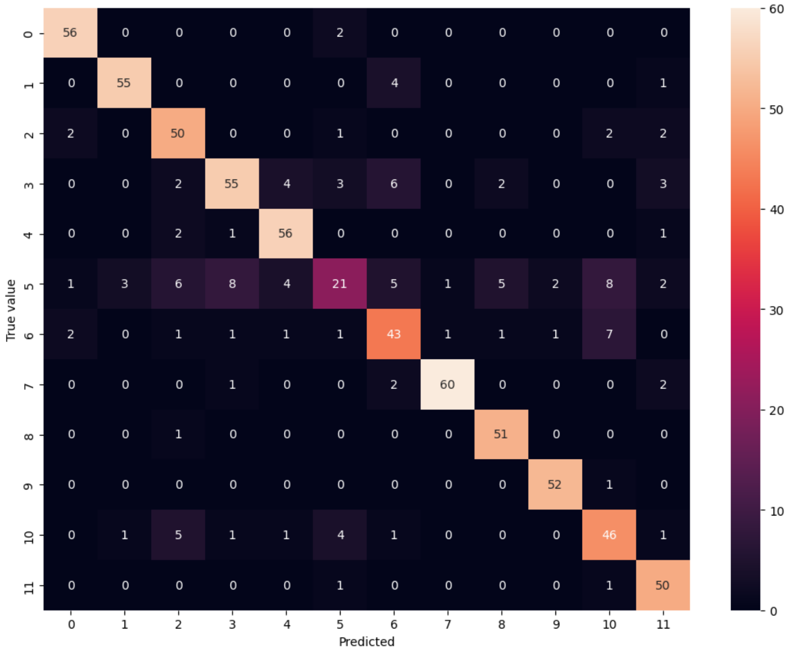
When observing the confusion matrix for the SVM [102] classifier it is clear a large proportion of the confusion came from classifying the depression disorder, as every cell off the diagonal for the true value of depression contained a value indicating a repeat error. This repeat pattern could be speculated that either depression shares a lot of commonalities with other disorders or that the data set is not sufficient enough to reach high levels of discrimination with this particular disorder.



	0	1	2	3	4	5	6	7	8	9	10	11
0	58	0	0	0	0	0	0	0	0	0	0	0
1	0	60	0	0	0	0	0	0	0	0	0	0
2	0	0	46	1	1	3	0	2	2	0	2	0
3	0	0	1	60	5	8	0	0	0	0	1	0
4	0	0	0	1	58	1	0	0	0	0	0	0
5	5	3	9	7	3	24	7	1	1	1	4	1
6	0	0	0	1	1	2	48	0	1	2	4	0
7	0	0	2	0	0	0	0	62	0	1	0	0
8	0	0	0	0	0	1	0	0	48	0	3	0
9	0	0	0	0	0	0	0	0	2	50	1	0
10	1	2	5	1	0	4	1	0	0	4	42	0
11	0	1	0	0	0	1	0	0	0	0	0	50

Figure 11 - Over sampled hyper tuned SVM confusion matrix

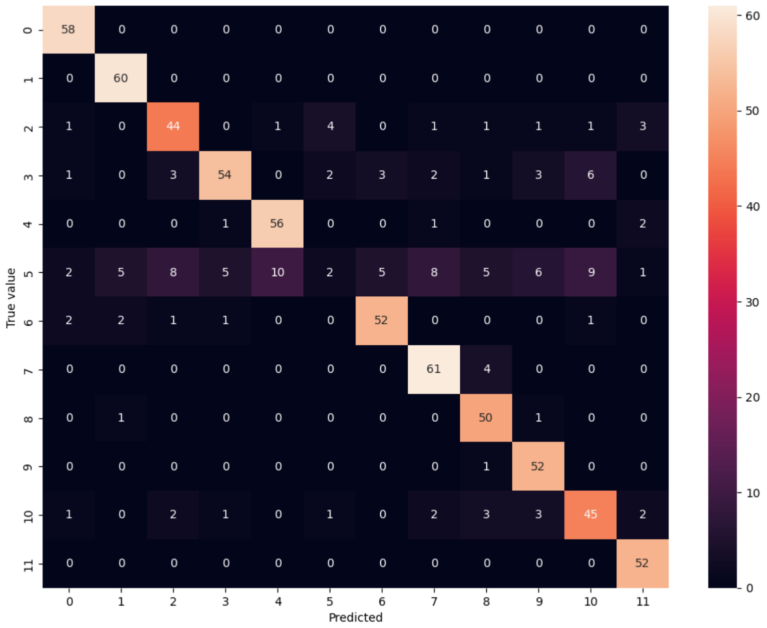
The repeat pattern for confusion with classification of the depression disorder was also seen in the Cat Boost [34] [94] [103] [93] confusion matrix [101] along with some additional confusion when assigning the healthy control which is definitely not optimal as you would never want to diagnose a patient with a treatment who doesn't require it as its both a waste of money and medication but also could harm the individual if they react badly to the treatment.



	0	1	2	3	4	5	6	7	8	9	10	11
0	56	0	0	0	0	2	0	0	0	0	0	0
1	0	55	0	0	0	0	4	0	0	0	0	1
2	2	0	50	0	0	1	0	0	0	0	2	2
3	0	0	2	55	4	3	6	0	2	0	0	3
4	0	0	2	1	56	0	0	0	0	0	0	1
5	1	3	6	8	4	21	5	1	5	2	8	2
6	2	0	1	1	1	1	43	1	1	1	7	0
7	0	0	0	1	0	0	2	60	0	0	0	2
8	0	0	1	0	0	0	0	0	51	0	0	0
9	0	0	0	0	0	0	0	0	0	52	1	0
10	0	1	5	1	1	4	1	0	0	0	46	1
11	0	0	0	0	0	1	0	0	0	0	1	50

Figure 12 - Over sampled hyper tuned Cat Boost confusion matrix

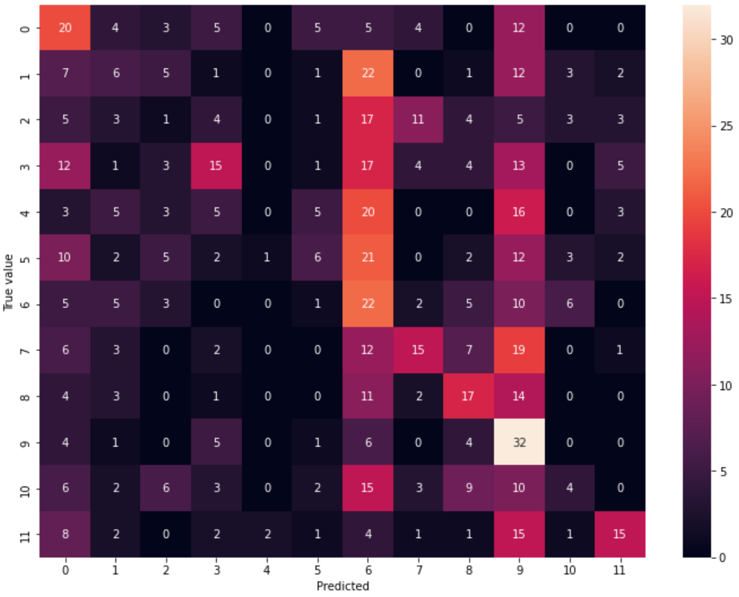
When observing the confusion matrix for the KNN classifier it was noted that depression was still the biggest weakness of the classifier. 10 labels were misclassified between depression and bipolar disorder indicating potential overlap in key distinguishing features for these 2 psychiatric disorders. To confirm the overlap more data would need supplied for training to determining the features that discriminate between them.



	0	1	2	3	4	5	6	7	8	9	10	11
0	58	0	0	0	0	0	0	0	0	0	0	0
1	0	60	0	0	0	0	0	0	0	0	0	0
2	1	0	44	0	1	4	0	1	1	1	1	3
3	1	0	3	54	0	2	3	2	1	3	6	0
4	0	0	0	1	56	0	0	1	0	0	0	2
5	2	5	8	5	10	2	5	8	5	6	9	1
6	2	2	1	1	0	0	52	0	0	0	1	0
7	0	0	0	0	0	0	0	61	4	0	0	0
8	0	1	0	0	0	0	0	0	50	1	0	0
9	0	0	0	0	0	0	0	0	1	52	0	0
10	1	0	2	1	0	1	0	2	3	3	45	2
11	0	0	0	0	0	0	0	0	0	0	0	52

Figure 13 - Over sampled hyper tuned KNN confusion matrix

To explore the models further via the confusion matrix it only made sense to observe a model with a much weaker performance such as Naïve bayes [97] [98] and the graphic below highlights how there is little to no consistency from the model when performing multi classification against the EEG data [13] as a majority of the grid squares contain some form of confusion/error.



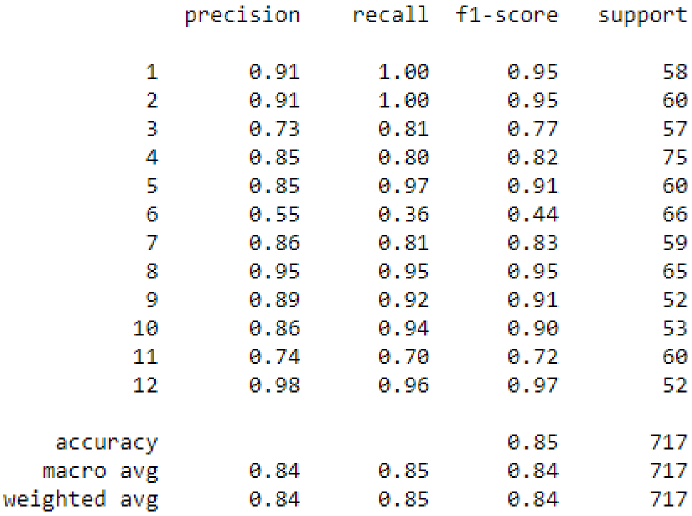
	0	1	2	3	4	5	6	7	8	9	10	11
0	20	4	3	5	0	5	5	4	0	12	0	0
1	7	6	5	1	0	1	22	0	1	12	3	2
2	5	3	1	4	0	1	17	11	4	5	3	3
3	12	1	3	15	0	1	17	4	4	13	0	5
4	3	5	3	5	0	5	20	0	0	16	0	3
5	10	2	5	2	1	6	21	0	2	12	3	2
6	5	5	3	0	0	1	22	2	5	10	6	0
7	6	3	0	2	0	0	12	15	7	19	0	1
8	4	3	0	1	0	0	11	2	17	14	0	0
9	4	1	0	5	0	1	6	0	4	32	0	0
10	6	2	6	3	0	2	15	3	9	10	4	0
11	8	2	0	2	2	1	4	1	1	15	1	15

Figure 14 - Over sampled hyper tuned Naive Bayes confusion matrix

Classification report analysis

The classification report [104] consists of a precision, recall and f1-score which all measure the model's accuracy. Precision measures the number of positive instances that were correctly identified, recall measures the proportion of actual instances correctly identified and f1-score is the mean of the precision and recall scores. The number values on the left side represent each of the disorders classified meaning observations can be made further on the classifier's strengths and weaknesses.

When observing the SVM [102] classification report Figure 15 [104] further emphasis on how the model had difficulty classifying depression disorder against others was displayed and it has impacted the classification of alcohol use disorder and schizophrenia. All other disorders achieved a respectable level of accuracy in the report reflecting in an above average model.



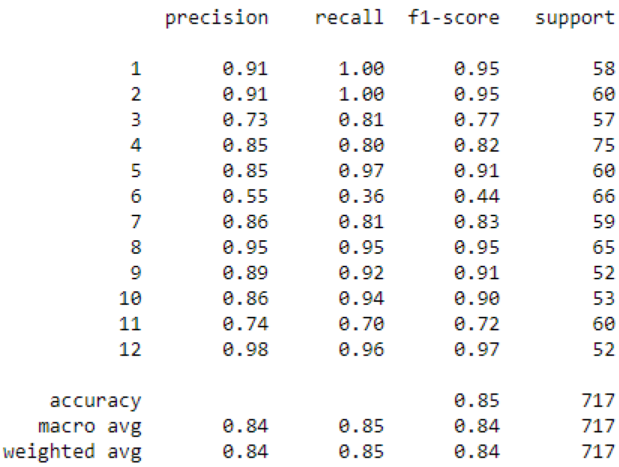
```
              precision    recall  f1-score   support

           1       0.91      1.00      0.95        58
           2       0.91      1.00      0.95        60
           3       0.73      0.81      0.77        57
           4       0.85      0.80      0.82        75
           5       0.85      0.97      0.91        60
           6       0.55      0.36      0.44        66
           7       0.86      0.81      0.83        59
           8       0.95      0.95      0.95        65
           9       0.89      0.92      0.91        52
          10       0.86      0.94      0.90        53
          11       0.74      0.70      0.72        60
          12       0.98      0.96      0.97        52

    accuracy                           0.85       717
   macro avg       0.84      0.85      0.84       717
weighted avg       0.84      0.85      0.84       717
```

Figure 15 - Over sampled hyper tuned SVM classification report

The Cat Boost [34] [94] [103] [93] classification report Figure 16 revealed a lot of similarities when compared to the SVM report in particular when the model attempts to classify the depression disorder the recall rate heavily dips to 36% with a precision of 55% reflecting in confusion when attempting to classify [13] EEG recordings of a subject with depression.



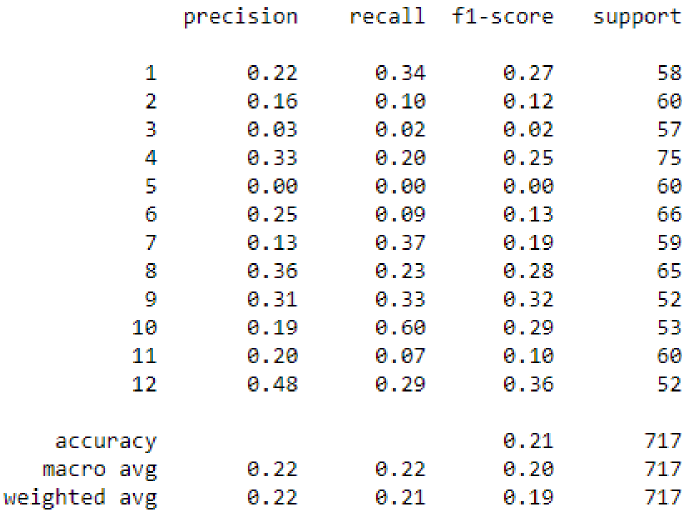
```
              precision    recall  f1-score   support

           1       0.91      1.00      0.95        58
           2       0.91      1.00      0.95        60
           3       0.73      0.81      0.77        57
           4       0.85      0.80      0.82        75
           5       0.85      0.97      0.91        60
           6       0.55      0.36      0.44        66
           7       0.86      0.81      0.83        59
           8       0.95      0.95      0.95        65
           9       0.89      0.92      0.91        52
          10       0.86      0.94      0.90        53
          11       0.74      0.70      0.72        60
          12       0.98      0.96      0.97        52

    accuracy                           0.85       717
   macro avg       0.84      0.85      0.84       717
weighted avg       0.84      0.85      0.84       717
```

Figure 16 - Over sampled hyper tuned Cat Boost classification report

The previous two classification reports [104] were of the best performing models, so to highlight the difference between the top and bottom end models a classification report for [97] [98] Naïve Bayes was examined. Naïve Bayes was the weakest classifier in terms of accuracy with consistently low results for precision, recall and f1-score. Social anxiety disorder achieved the highest accuracy in this report but was still below 50% meaning there would always be uncertainty if this model correctly classified.



```
              precision    recall  f1-score   support

           1       0.22      0.34      0.27        58
           2       0.16      0.10      0.12        60
           3       0.03      0.02      0.02        57
           4       0.33      0.20      0.25        75
           5       0.00      0.00      0.00        60
           6       0.25      0.09      0.13        66
           7       0.13      0.37      0.19        59
           8       0.36      0.23      0.28        65
           9       0.31      0.33      0.32        52
          10       0.19      0.60      0.29        53
          11       0.20      0.07      0.10        60
          12       0.48      0.29      0.36        52

    accuracy                           0.21       717
   macro avg       0.22      0.22      0.20       717
weighted avg       0.22      0.21      0.19       717
```

Figure 17 -Over sampled hyper tuned Naive Bayes classification report

Proof of concept [105]

Requirements for POC

POC is an alternative faster approach to achieve funding, it is made up of evidence obtained from either research or the demonstration of a project. The POC leads to a precommercial grant awarding fund to help entrepreneurs in Northern Ireland (run by Catalyst) [106]. To qualify for the POC grant with Catalyst the project must show evidence of commercial value, a target market must be demonstrated and a demonstration providing further evidence of a model that can prove substantial precision.

Analysis to determine if POC was met

The project does meet the POC requirements set by Catalyst as mentioned above in the supporting literature the potential to impact the pharmaceutical market is estimated to be in the £ billions, the selected market would be those suffering from a psychiatric related disorder and the models promise of above chance accuracy with a sufficient balanced data would satisfy Catalyst requirements.

Evaluation

Verification and Validation

The project was verified with an independent coordinator from Ulster University [107]before proceeding, during discussions with the coordinator a predefined set of aims and objectives were set. The independent also verified the dataset as a link was provided to the website along with personal validation via the preprocessing stage to determine that data was fit for purpose. It was during this stage it was noted that the data did not consist of raw EEG [13] recordings, but of pre-processed extracted means of each channel per band. This was a setback as ideally the raw EEG [13] would have been obtained so that demonstrations of artifact removal could be performed along with processing via DL algorithms. Adjustments were made to instead process the recordings through a range of popular ML algorithms as the solution to handling the dataset.

Results from the variations of multiple classifiers were compared and evaluated against the initial aims and objectives. This helped to determine the strengths and weaknesses of the research and classifiers developed. The selected research methodology and classifiers also reflected industry standard processes and concepts.

In regards to the methodology used to develop this project a business understanding was laid out as the project mentions not only how there is commercial opportunity to impact the pharmaceutical market, but also laid out the current problems with growth in psychiatric disorders and the availability of well labelled public EEG datasets.

During the data understanding stage of the project life cycle the dataset was selected according to the aims and objectives of the project. With the selected dataset from a reliable source acquired it was then put through a range of visualisation techniques to understand the data and determine if there were missing entries, irrelevant data or if the dataset was unbalanced which it was later determined it was.

Next the data preparation stage was used to remove irrelevant data which included the fields, 'gender', 'age', 'date of EEG recording', 'education', 'IQ' and 'disorder'. This meant only the specific disorder along with the PSD and brain wave channel coherence EEG recordings were left.

The modelling stage is where all of the various architectures were applied to the EEG data, the initial base algorithms were applied, under sampling was applied, over sampling and over sampling with hyper parameter tuning with cross-validation was applied to achieve the best results in terms of accuracy and recall from the SVM, Cat Boost and KNN classifiers.

Evaluation of the models were performed with the score method from scikit learn, confusion matrices were used to find the models strengths and weaknesses and lastly a classification report was used for each model to evaluate their recall and precision for each specific disorder.

Deployment of the algorithm in this project did not take place as the model with its current dataset is not fit for generalisation. A POC from this project would be used to raise funds to run a trial to record an in-house balanced dataset so that a model fit for generalisation could possibly be developed. It would also have preferred the dataset to contain the raw EEG recordings as this is the authentic raw recording of the patient.

Maintenance in this project was voided as previously mentioned the model in my opinion was not fit for generalisation and therefore was not deployed.

Critical Appraisal

The main objective was to develop a POC to determine if it is possible to perform multi classification on EEG [13] data to diagnose a range of psychiatric disorders and achieve an accuracy of 95% or above whilst exploring other architectures separate from the publication [18] upon which laid the foundation for this research. Investigating a range of classifiers was necessary to extract any form of relationships or differentials in the data that could help with the classification and therefore the diagnosis of a patient.

Pharmaceutical market growth in this sector indicates either our current treatments are not working effectively, mis-diagnosis could be taking place so an increased understanding of the disorder is required or that even potentially pharmaceutical companies are seeking profit versus curing their patients as this would mean one less customer to their treatment. It is with these speculations that morally society has an obligation to help people suffering with a range of psychiatric disorders and the first step in this process is to be able to develop a deep understanding of the disorders in order to improve or design more effective treatments.

SVM [102] architecture supported the high dimensional requirements for processing EEG [13] data hence its effective performance whilst Cat Boost [34] [94] [103] [93] was successful in achieving high levels of accuracy for category-based classification which complimented the dataset as this was a multi class classification problem. KNN's performance surprised myself as public literature did not explore the use of a much simpler classifier and hence it may have been overlooked.

The project does meet the requirements to act as a POC to help acknowledge how EEG [13] data can be used to diagnose a range of disorders along with most effective architectures, but with the current data set used it is not possible to say the models manufactured are fit for generalisation as the data was imbalanced, medications were not controlled, recordings were only of locals and there is not a sufficient quantity of recorded psychiatric disorders to represent a global audience. The project achieved a peak accuracy score of 86.77% with the SVM classifier this reflected in an above chance accuracy, but it did not meet the objective as an accuracy score of 95% was desired.

References

[1] BBC, "Nearly half a million more adults on antidepressants in England," [Online]. Available: https://www.bbc.co.uk/news/health-62094744. [Accessed 2023 08 19].

[2] Business Research Comany, "Antidepressants Global Market Report," [Online]. Available: https://www.thebusinessresearchcompany.com/report/antidepressant-global-market-report. [Accessed 11 08 2023].

[3] UN News, "Nearly one billion people have a mental disorder: WHO," [Online]. Available: https://news.un.org/en/story/2022/06/1120682.

[4] GHDx, "GHDx," [Online]. Available: https://vizhub.healthdata.org/gbd-results/.

[5] Research and markets, "Psychiatrist Global Market Report," [Online]. Available: https://www.researchandmarkets.com/report/psychiatry.

[6] Gov.uk, "Premature mortality in adults with severe mental illness (SMI)," [Online]. Available: https://www.gov.uk/government/publications/premature-mortality-in-adults-with-severe-mental-illness/premature-mortality-in-adults-with-severe-mental-illness-smi.

[7] M. Luciano, "Mortality of people with severe mental illness: Causes and ways of its reduction.," [Online]. Available: https://www.frontiersin.org/research-topics/17988/mortality-of-people-with-severe-mental-illness-causes-and-ways-of-its-reduction.

[8] C. C. &. Consulting, "HOW SELF-ESTEEM RELATES TO PHYSICAL & MENTAL HEALTH," [Online]. Available: https://www.cognition.org.uk/coaching-hypnotherapy/blog/how-self-esteem-relates-to-physical-mental-health#:~:text=Self-esteem%20is%20related%20to,poor%20physical%20and%20mental%20health.

[9] A. S. a. M. c. H. Texas, "Depression Diagnosis and Drug Response," [Online]. Available: https://arxiv.org/ftp/arxiv/papers/2303/2303.06033.pdf.

[10] IBM, "Recurrent Neural Network," [Online]. Available: https://www.ibm.com/topics/recurrent-neural-networks.

[11] IBM, "Convolutional neural networks," [Online]. Available: https://www.ibm.com/topics/convolutional-neural-networks.

[12] MathWorks, "Long short term memory networks," [Online]. Available: https://www.mathworks.com/help/deeplearning/ug/long-short-term-memory-networks.html.

[13] Mayoclinic, "EEG explained," [Online]. Available: https://www.mayoclinic.org/tests-procedures/eeg/about/pac-20393875.

[14] S. kumar, "Prediction of Depression from EEG Signal Using Long Short Term Memory(LSTM)," [Online]. Available: https://ieeexplore.ieee.org/document/8862560.

[15] Research Gate, "Transformer vs LSTM methods," [Online]. Available: https://www.researchgate.net/figure/Transformer-based-VS-LSTM-based-models-performance-comparison-with-different_fig2_361098572#:~:text=Accuracies%20of%20transformer%2Dbased%20models,accuracies%20of%20LSTM%2Dbased%20models.&text=Integrating%20machine%20le.

[16] J. K. Johannesen, "Machine learning identification of EEG features predicting working memory performance in schizophrenia and healthy adults," [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/27375854/.

[17] scikit learn, "SVM," [Online]. Available: https://scikit-learn.org/stable/modules/svm.html.

[18] S. S. A. A. M. A. M. Y. a. A. S. M. Wajid Mumtaz, "A machine learning framework involving EEG-based functional connectivity to diagnose major depressive disorder (MDD)," [Online]. Available: https://link.springer.com/article/10.1007/s11517-017-1685-z.

[19] Modma.lzu.edu.cn, "EEG data source," [Online]. Available: http://modma.lzu.edu.cn/data/application/#data_1.

[20] World health organisation, "Mental Disorders," [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders. [Accessed 20 08 2023].

[21] P. Dave, "Neural Conduction, Action Potential and Synaptic transmission," [Online]. Available: https://www.youtube.com/watch?v=zHJ3h675nNk&list=LL&index=46&t=3s.

[22] IBM, "Machine Learning Vs Deep learning Neural Networks," [Online]. Available: https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/.

[23] MDPI, "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology," [Online]. Available: https://www.mdpi.com/2504-4990/3/2/392.

[24] S. Studer, "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology," [Online]. Available: https://www.mdpi.com/2504-4990/3/2/392.

[25] R. Wirth, "CRISP-DM," [Online]. Available: https://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf.

[26] Python, "Python.org," [Online]. Available: https://www.python.org/.

[27] JetBrains, "JetBrains Pycharm," [Online]. Available: https://www.jetbrains.com/pycharm/.

[28] Microsoft, "Microsoft Excel," [Online]. Available: https://www.microsoft.com/en-gb/microsoft-365/excel.

[29] FileFormat, "CSV," [Online]. Available: https://docs.fileformat.com/spreadsheet/csv/#:~:text=csv%20(Comma%20Separated%20Values)%20extension,one%20storage%20system%20to%20another.

[30] Pandas, "Pandas DataFrame," [Online]. Available: https://pandas.pydata.org/.

[31] Numpy.org, "Numpy," [Online]. Available: https://numpy.org/.

[32] matplotlib.org, "MatPlotLib," [Online]. Available: https://matplotlib.org/.

[33] SeaBorn,org, "SeanBorn," [Online]. Available: https://seaborn.pydata.org/.

[34] Catboost, "Catboost," [Online]. Available: https://catboost.ai/.

[35] Scikit learn, "K nearest neighbours," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html.

[36] lightgbm, "LightGBM," [Online]. Available: https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.html.

[37] sci kit learn, "Linear Regression," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.

[38] IBM, "Naive Bayes," [Online]. Available: https://www.ibm.com/docs/en/ias?topic=procedures-naive-bayes.

[39] Scikit learn, "Random Forest Classififer," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[40] Scikit learn, "Decision Trees," [Online]. Available: https://scikit-learn.org/stable/modules/tree.html.

[41] XGBoost, "XGBoost," [Online]. Available: https://xgboost.readthedocs.io/en/stable/parameter.html.

[42] Sci-kit learn, "RepeatedKFold," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RepeatedKFold.html.

[43] Matplotlib, "Matplotlib," [Online]. Available: https://matplotlib.org/3.5.3/api/_as_gen/matplotlib.pyplot.html.

[44] Adobe, "Photopshop," [Online]. Available: https://www.adobe.com/uk/creativecloud/business/teams.html?sdid=6S3T725T&mv=search&ef_id=Cj0KCQjwxuCnBhDLARIsAB-cq1oFb4NR7EbV9bJGiK9VQ5xj_NYEJn_wcpskINp2qVCFuk_3Rnt0J6UaAi-mEALw_wcB:G:s&s_kwcid=AL!3085!3!655751656058!e!!g!!photoshop!20010189115!1478448059. [Accessed 06 09 2023].

[45] NHS, "Electroencephalogram (EEG)," [Online]. Available: https://www.nhs.uk/conditions/electroencephalogram/.

[46] QBI.UQ.EDU, "What is a neron and the components?," [Online]. Available: https://qbi.uq.edu.au/brain/brain-anatomy/what-neuron#:~:text=A%20dendrite%20(tree%20branch)%20is,structures%20on%20them%20called%20spines..

[47] Kaggle, "EEG data of pyschiatric disorders dataset," [Online]. Available: https://www.kaggle.com/datasets/shashwatwork/eeg-psychiatric-disorders-dataset.

[48] OSF Home, "Su Mi Park," [Online]. Available: https://osf.io/pm3yt/.

[49] ORCid, "ORCid," [Online]. Available: https://orcid.org/0000-0001-5200-9507.

[50] L. Lakshmanan, "What is a feature / weight?," [Online]. Available: https://towardsdatascience.com/why-feature-weights-in-a-machine-learning-model-are-meaningless-b0cd22a4c159.

[51] Neuroscan, "NeuroScan," [Online]. Available: https://compumedicsneuroscan.com/applications/eeg/.

[52] National Library of Medicine, "Common Artifacts During EEG Recording," [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK390358/.

[53] Codecademy, "IDE," [Online]. Available: https://www.codecademy.com/article/what-is-an-ide.

[54] SIEMENS, "Power Spectral Density," [Online]. Available: https://community.sw.siemens.com/s/article/what-is-a-power-spectral-density-psd.

[55] National Institute of Mental Health, "Post Traumatic Stress Disorder PTSD," [Online]. Available: https://www.nimh.nih.gov/health/topics/post-traumatic-stress-disorder-

ptsd#:~:text=Post%2Dtraumatic%20stress%20disorder%20(PTSD)%20is%20a%20disorder%20that,or%20respond%20to%20potential%20danger.

[56] National Institute of Mental Health, "Schizophrenia," [Online]. Available: https://www.nimh.nih.gov/health/topics/schizophrenia#:~:text=What%20is%20schizophrenia%3F,for%20their%20family%20and%20friends.

[57] NHS, "Depression In adults," [Online]. Available: https://www.nhs.uk/mental-health/conditions/depression-in-adults/overview/.

[58] NHS, "Social Anxietyt Disorder," [Online]. Available: https://www.nhs.uk/mental-health/conditions/social-anxiety/.

[59] National Institute of Mental Health, "BiPolar," [Online]. Available: https://www.nimh.nih.gov/health/publications/bipolar-disorder.

[60] National Institute of Mental Health, "Obsessive Compulsive Disorder," [Online]. Available: https://www.nimh.nih.gov/health/topics/obsessive-compulsive-disorder-ocd#:~:text=Obsessive%2Dcompulsive%20disorder%20(OCD),to%20repeat%20over%20and%20over.

[61] National Institute on Alcohol Abuse and Alcoholism, "Alcohol Use Disorder," [Online]. Available: https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/understanding-alcohol-use-disorder.

[62] NHS, "Panic Disorder," [Online]. Available: https://www.nhsinform.scot/illnesses-and-conditions/mental-health/panic-disorder#:~:text=Panic%20disorder%20is%20where%20you,to%20stressful%20or%20dangerous%20situations.

[63] National Library of Medicine, "Adjustment Disorder," [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3701359/.

[64] Centers for Disease Control and Prevention, "Behavior Disorder," [Online]. Available: https://www.cdc.gov/childrensmentalhealth/behavior.html.

[65] National Library of Medicine, "Acute Stress Disorder," [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK560815/.

[66] imbalanced.org, "Random Under Sampler / Imbalanced dataset," [Online]. Available: https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html.

[67] NeuroSky, "NeuroSky explaing waves," [Online]. Available: https://neurosky.com/2015/05/greek-alphabet-soup-making-sense-of-eeg-bands/#:~:text=In%20our%20previous%20blog%2C%20we,waveform%20captured%20at%20an%20electrode.

[68] Research Gate , "10-20 Electrode Placement rules," [Online]. Available: https://www.researchgate.net/figure/The-placement-positions-of-EOG-Fp1-and-Fp2-electrodes-E1-and-E2-are-the-left-and-right_fig1_335074690.

[69] S. M. O. S. M. D. H. &. J. V. M. Steven X. Moffett, "Dynamics of high frequency brain activity," [Online]. Available: https://www.nature.com/articles/s41598-017-15966-6.

[70] S. Tavasoli, "Simon Tavasoli," [Online]. Available: https://www.linkedin.com/in/simontavasoli/?originalSubdomain=ca.

[71] S. Gupta, "Pros and cons of various Machine Learning algorithms," [Online]. Available: https://towardsdatascience.com/pros-and-cons-of-various-classification-ml-algorithms-3b5bfb3c87d6.

[72] A. Saeedi1, "Depression Diagnosis and Drug Response," [Online]. Available: https://arxiv.org/ftp/arxiv/papers/2303/2303.06033.pdf.

[73] IBM, "What is logistic regression?," [Online]. Available: https://www.ibm.com/topics/logistic-regression.

[74] IBM, "What is a Decision Tree?," [Online]. Available: https://www.ibm.com/topics/decision-trees.

[75] IBM, "What is random forest?," [Online]. Available: https://www.ibm.com/topics/random-forest.

[76] IBM, "What is the k-nearest neighbors algorithm?," [Online]. Available: https://www.ibm.com/topics/knn.

[77] Python is a planet, "Pros and Cons of Unsupervised Learning," [Online]. Available: https://pythonistaplanet.com/pros-and-cons-of-unsupervised-learning/?utm_content=cmp-true.

[78] N. Khandelwal, "Introduction to Semi-supervised Learning," [Online]. Available: https://www.shiksha.com/online-courses/articles/introduction-to-semi-supervised-learning/.

[79] Pythonisaplanet, "Pros and Cons of Supervised Machine Learning," [Online]. Available: https://pythonistaplanet.com/pros-and-cons-of-supervised-machine-learning/.

[80] IBM, "What is supervised learning?," [Online]. Available: https://www.ibm.com/topics/supervised-learning.

[81] C. Leschinski, "Are you interpreting your logistic regression correctly?," [Online]. Available: https://towardsdatascience.com/are-you-interpreting-your-logistic-regression-correctly-d041f7acf8c7#:~:text=Logistic%20regression%20was%20introduced%20in,the%20first%20model%20we%20try.

[82] Opengenus.org, "Advantages and Disadvantages of Logistic Regression," [Online]. Available: https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/.

[83] AI and the LinkedIn community, "What are the benefits of elastic net?," [Online]. Available: https://www.linkedin.com/advice/0/what-some-common-pitfalls-challenges-elastic#:~:text=One%20of%20the%20benefits%20of,select%20one%20predictor%20over%20another.

[84] C. D. Mol, "Elastic-Net Regularization in Learning Theory," [Online]. Available: https://www.researchgate.net/publication/222674725_Elastic-Net_Regularization_in_Learning_Theory.

[85] J. Brownlee, "How to Develop Elastic Net Regression Models in Python," [Online]. Available: https://machinelearningmastery.com/elastic-net-regression-in-python/.

[86] History Of datascience, "Decision Tree and Random Forest Algorithms: Decision Drivers," [Online]. Available: https://www.historyofdatascience.com/decision-tree-and-random-forest-algorithms-decision-drivers/.

[87] D. Leventis, "XGBoost Mathematics Explained," [Online]. Available: https://dimleve.medium.com/xgboost-mathematics-explained-58262530904a.

[88] V. Morde, "XGBoost Algorithm: Long May She Reign!," [Online]. Available: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d.

[89] Simplilearn, "What is XGBoost? An Introduction to XGBoost Algorithm in Machine Learning," [Online]. Available: https://www.simplilearn.com/what-is-xgboost-algorithm-in-machine-learning-article.

[90] G. Ke1, "LightGBM: A Highly Efficient Gradient Boosting," [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[91] P. Mandot, "What is LightGBM, How to implement it? How to fine tune the parameters?," [Online]. Available: https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc.

[92] S. SURANA, "What is Light GBM? Advantages & Disadvantages? Light GBM vs XGBoost?," [Online]. Available: https://www.kaggle.com/discussions/general/264327.

[93] M. Yawar, "CatBoost-ML," [Online]. Available: https://www.codingninjas.com/studio/library/catboost-ml.

[94] A. Oppermann, "What Is CatBoost?," [Online]. Available: https://builtin.com/machine-learning/catboost.

[95] R. Gandhi, "Support Vector Machine — Introduction to Machine Learning Algorithms," [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47#:~:text=Support%20vectors%20are%20data%20points,help%20us%20build%20our%20SVM.

[96] K. Pykes, "Random Forest Overview," [Online]. Available: https://towardsdatascience.com/random-forest-overview-746e7983316#:~:text=The%20Random%20forest%20is%20an,by%20Leo%20Breiman%20in%202001%C2%B2.

[97] IBM, "What are Naïve Bayes classifiers?," [Online]. Available: https://www.ibm.com/topics/naive-bayes.

[98] Upgrad, "Naive Bayes Explained," [Online]. Available: https://www.upgrad.com/blog/gaussian-naive-bayes/.

[99] Z. Zhang, "Naive Bayes Explained," [Online]. Available: https://towardsdatascience.com/naive-bayes-explained-9d2b96f4a9c0.

[100] W. ROSINSKI, "SHAP API feature importance," [Online]. Available: https://www.kaggle.com/code/wrosinski/shap-feature-importance-with-feature-engineering.

[101] scikit learn, "Confusion Matrix," [Online]. Available: https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html.

[102] scikit learn, "SVM," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

[103] M. Raval, "Understanding CatBoost Algorithm," [Online]. Available: https://medium.com/analytics-vidhya/catboost-101-fb2fdc3398f3.

[104] Muthukrishnan, "Classification Report explained," [Online]. Available: https://muthu.co/understanding-the-classification-report-in-sklearn/.

[105] EduGrowth, "Proof of Concept POC," [Online]. Available: https://edugrowth.org.au/2022/04/13/proof-of-concept-catalyst-to-new-market-entry/.

[106] Catalyst, "Catalyst," [Online]. Available: https://www.youtube.com/watch?v=Sjf6JRjeX1A.

[107] Ulster University, "Ulster University," [Online]. Available: https://www.ulster.ac.uk/.