Psychiatry and Machine Learning: Unlocking New Frontiers in Mental Health Diagnosis and Treatment

Atharva Tiwari

May 2024

WPI

Contents

1 Abstract **3**

2 Introduction **4**

3 Related Work **5**

4 Methods **8**

4.1 Data Cleaning and Joining 8

4.2 Exploratory Data Analysis (EDA) 8

5 Machine Learning Model Selection and Results **14**

5.1 Support Vector Machine (SVM) 14

5.2 Random Forrest Classifier 16

5.3 Convolutional Neural Networks (CNN) 18

6 Conclusion **20**

List of Figures

1 *Electrode placement in EEG (a) Location and nomenclature of the scalp electrodes in the international 10-20 system as standardized by the International Federation of Clinical Physiology (IFCN), (b) Location and nomenclature of the scalp electrodes.* 4

2 *Histogram of Main Disorders from EEG Psychiatric Disorders Dataset* 8

3 *Results of extracting important features using Gini Importance* 10

4 *Histogram of the most important features* 11

5 *Box-plots of the most important features* 12

6 *Seaborn Correlation Matrix for most important features* 13

7 *Classification Report from best run of SVM algorithm* 15

8 *Confusion Matrix showcasing the accuracy of classification.* 15

9 *Classification Report from the best run of the Random Forrest Classifier* 16

10 *Confusion Matrix showing the accuracy of the Random Forrest Classifier* 17

11 *Precision-Recall Curve for the Random Forrest Classifier Model* 18

12 *Results from the CNN showing the layers as well as the Test Loss and Accuracy* 19

1 Abstract

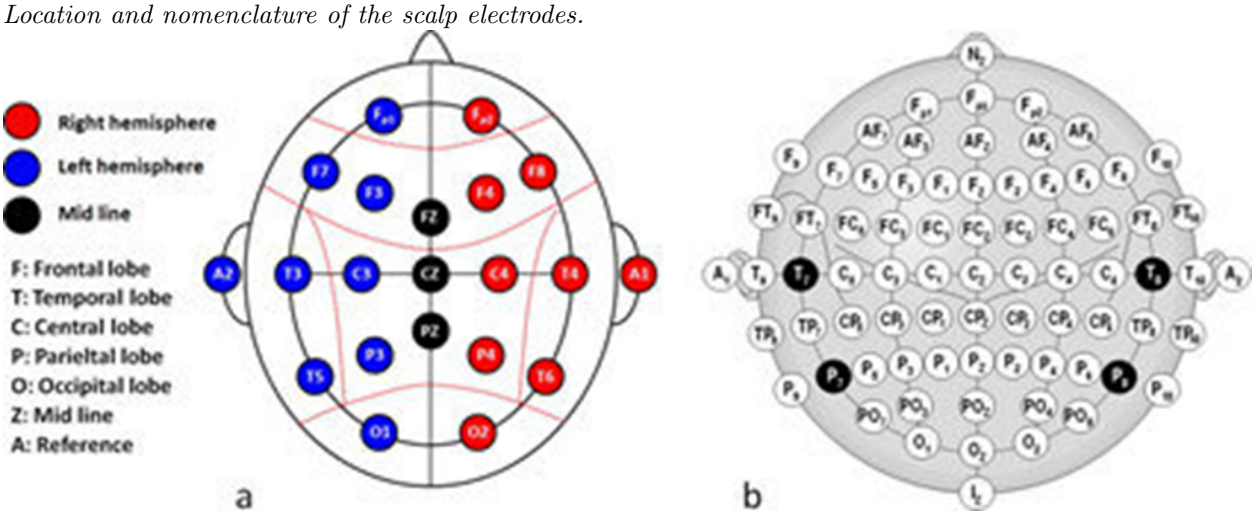
This paper was the culmination of research on psychiatric disorder, electroencephalogram (EEG) data and machine learning models for classification. This research paper examines the role that Machine Learning (ML) can play in aiding psychiatrists in the diagnoses of psychiatric disorder. This paper discusses why an ML model would be helpful in the current mental health climate and to prevent misdiagnoses. It discusses the approach taken to parse and clean the data, analyze it and prepare it for machine learning analysis. Finally, this cleaned data is used in Support Vector Machines (SVM), Random Forest, and CNNs to see which algorithm yields the best model. This paper concludes with a discussion of these approaches and why a multi-class classification model cannot be achieved with this data.

2 Introduction

According to the World Health Organization (WHO) in 2020 there were 970 million people globally who were living with a mental disorder (WHO). This statistic highlights how many people struggle worldwide but fails to capture their true struggle. However, the journey to getting better first involves seeking professional care and help. When patients first arrive either at a hospital or a psychiatrist's office, they are asked a series of questions and certain tests are conducted to diagnose the patient. However, in the medical field the diagnosis of mental health disorders can often be wrong leading to incorrect treatment, confusion, and more mental trauma for the patient Akers (2023). In a recent study performed by doctors in Canada they uncovered that within their study "misdiagnosis rates reached 65.9% for major depressive disorder, 92.7% for bipolar disorder, 85.8% for panic disorder, 71.0% for generalized anxiety disorder, and 97.8% for social anxiety disorder" Vermani et al. (2011) . The authors had gone onto conclude that with poor detection for certain disorders an aid for psychiatrists can be beneficial. This is where EEG reading can help!

In a recent study performed by a group of psychiatric researchers in South Korea used resting state EEG for diagnosing psychiatric disorders. Their results were compiled into a dataset titled *EEG Psychiatric Disorders Dataset*. This dataset contains detailed information about the 945 patients that participated in the study ranging from their age, main diagnosis, specific diagnosis, and their education level. The remainder of their data is EEG readings from separate locations of the brain as seen in the figure below.

Figure 1: *Electrode placement in EEG (a) Location and nomenclature of the scalp electrodes in the international 10-20 system as standardized by the International Federation of Clinical Physiology (IFCN), (b) Location and nomenclature of the scalp electrodes.*



Note. From "Noninvasive Electromagnetic Methods for Brain Monitoring: A Technical Review" by Tushar Kanti Bera (2015). Intelligent Systems, 10.1007/978-3-319-10978-7_3

3 Related Work

As Machine Learning (ML) and Artificial Intelligence (AI) grow more popular in the world with many disciplines using the technology to improve their competitive edge. In the medical world this technology is being used for early cancer detection, improving Brain Computer Interfaces (BCI) and in medical imaging. Despite these popular uses, many are unaware of how this technology is being used in psychiatry to improve the diagnosis, treatments and outcomes of those diagnosed with psychiatric disorders.

Firstly, Chen et al. discuss how Modern machine learning approaches can be used in psychiatry and how recent advancements in ML and AI have enabled properly engineered algorithms to be able to "analyze complex patterns of neural and behavioral data for psychiatry" Chen et al. (2022). This paper delves into a discussion of various machine learning algorithms and their use case in psychiatry. For example, in this paper the authors state how Unsupervised Learning models such as K-means clustering was used with rs-EEG data, and the researchers discovered that there were two trans-diagnostic sub-types were identified and this was used to determine how different individuals respond differently to psychotherapy and anti-depressant medication. However, later on in the paper the authors discuss that different models may be required to aid in diagnosis because one method may not capture the complexities of the data and may lead to misdiagnosis, this is why the authors propose that ML should be applied to all parts of the patients journey. Based on this analysis I postulated whether EEG data can be used to classify individuals into clusters correlating to a psychiatric disorder. This research provided me with some possible models that I can use to generate my models. The table below shows Categories of ML, concepts, typical methods, and their representative applications.

Table 1. Categories of ML, concepts, typical methods, and their representative applications

Learning category	Concepts	Representative methods	Applications
Supervised	learning from labeled data to predict class/clinical measures	SVM, random forest, sparse learning, ensemble learning	Disease diagnosis, prognosis, treatment outcome prediction
Unsupervised	learning from unlabeled data to uncover structure and identify subgroups	Hierarchical clustering, K-means, PCA, CCA	Disease subtyping, normative modeling, identify behavioral and neurobiological dimension
Semi-supervised	learning from both labeled and unlabeled data to perform supervised or unsupervised tasks	multi-view learning, Laplacian regularization, semi-supervised clustering	multi-modal analysis, joint disease subtyping and diagnosis, prediction with incomplete data
Deep	learning hierarchies and non-linear mappings of features for higher-level representations, can be either supervised or unsupervised	CNN, deep autoencoder, GCN, RNN, LSTM, GAN	a large class of generic learning problems
Reinforcement	solving temporal credit assignment problems, optimal control, trial-and-error learning	temporal difference learning, Q-learning, actor-critic model, dynamic programming	online control, modeling of decision-making and choiced behaviors

Note. From "Modern views of machine learning for precision psychiatry" by Zhe Sage Chen et al. (2022). ScienceDirect, 10.1016

Next, in an article that discusses Machine Learning approaches for Clinical Psychology and Psychiatry, Dwyer et al discuss how the advent of machine learning algorithms are used in our daily lives and are becoming more prevalent in health care. The authors then discuss how there are commercial companies that are using Machine Learning algorithms to aid in psychiatric diagnoses. Subsequently, a discussion on the types of psychiatric problems that machine learning can be used to are discussed, "the four main categories are diagnosis, prognosis, treatment prediction, and the detection and monitoring of potential biomarkers" Dwyer et al. (2018). Among the discussion of these problems and how machine learning that can be applied to these problems one important fact that kept on being brought up was the need for these models to be accurate and be generalized to be integrated to clinical care. This raised some concern with regards to the data set that I am dealing with because, as seen in the Exploratory Data Analysis (EDA) below the data was not very diverse. Due to this reason I am concerned as if my model, if generated, will be able to scale to a clinical application. Another important discussion in this paper was with regards to Feature Selection for machine learning models, this is because with EEG data there is a lot of data and it is important to be

able to extract the most important features so that "over fitting is reduced and generalizability is increased" (Dwyer et al., 2018). Lastly, another important discussion some of the reasons that are holding back machine learning models for psychiatry. One of the reasons was hat there is not enough diverse data and as a result of this the models that are generated are unable to be scaled to a clinical application.

4 Methods

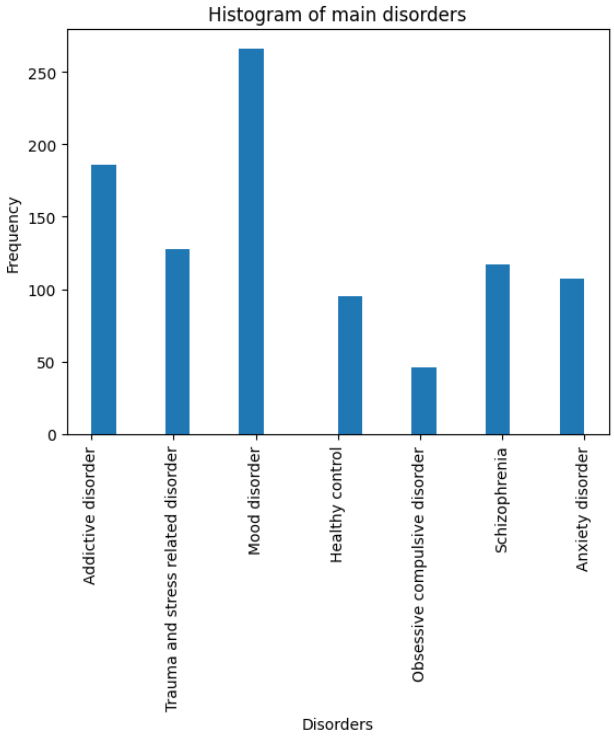
4.1 Data Cleaning and Joining

When conducting some initial data analysis there were occasions where there was some data that was not present indicated by a "NaN" in the pandas data frame. When these were encountered these values were dropped rather than being set to 0 as these were for EEG data and if the column value was set to 0 this could throw off the data that was present in the column. Furthermore, there were many columns that were present in the data, about 1152 columns per patient. This is a lot of data and further research had indicated that it would be best to calculate feature importance and use the ten most important features going forward.

4.2 Exploratory Data Analysis (EDA)

In this section I will be diving into statistics for the dataset. First, I look into what the unique disorders there are and witness how the data distribution is like within the dataset. As seen in the visual below it can be seen that the bulk of the data is concentrated in the Mood Disorder and the Addictive Disorder. Due to this uneven spread in the data I suspected that this would impact the output of the model. As it turns out this was true, there will be more on this soon.

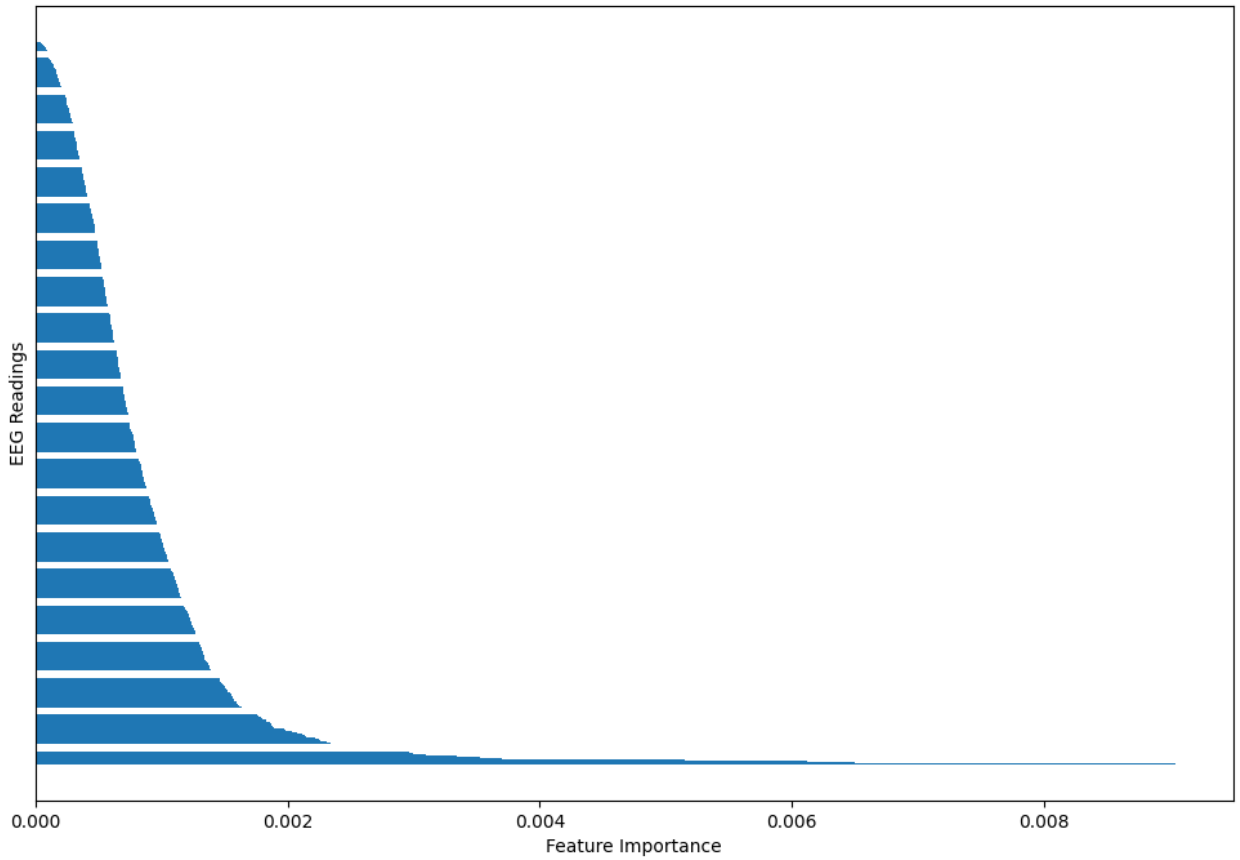
Figure 2: *Histogram of Main Disorders from EEG Psychiatric Disorders Dataset*



Furthermore, there were many columns that were present in the data, about 1152 columns per patient. This is a lot of data and further research had indicated that it would be best to calculate feature importance and use the ten most important features going forward. *Why is Feature Importance so valuable? Selecting important features in machine learning is valuable for several reasons. Firstly, it helps to reduce the dimensions of the data, which can lead to simpler, more interpret-able models and faster training times, especially with large datasets. By focusing on the most relevant features, the model can capture the essential patterns in the data while ignoring noise or irrelevant information, thus improving its generalization performance on unseen data. Additionally, feature selection can enhance model transparency and explain-ability, as it highlights the factors driving the model's predictions, making it easier for stakeholders to understand and trust the model's decisions. Moreover, feature selection can also aid in addressing issues like over-fitting, where the model learns to memorize the training data rather than generalize to new instances, by promoting a more parsimonious representation of the underlying data relationships. Especially, with EEG data this can be important as some of the node location can throw off the output of the model when considering using it for psychiatric diagnoses. The approach that I used was using Gini Importance technique.

Now what is Gini Importance? The Gini Importance technique is a method used in machine learning to determine the importance of features when building predictive models, particularly in decision tree-based algorithms like random forests. It calculates the total decrease in node impurity (measured by the Gini index) caused by a particular feature across all decision trees in the forest. In essence, it assesses how much each feature contributes to the model's ability to make accurate predictions. Features with higher Gini Importance scores are considered more influential in the model's decision-making process, making them crucial for understanding the underlying relationships in the data and optimizing model performance. The results as seen below yielded ten important features upon which further statistical analysis was conducted.

Figure 3: *Results of extracting important features using Gini Importance*



In the histogram plot below I take the most important features that were retrieved from the feature importance algorithm above and plot them. This uni-variate analysis shows for the EEG data the range of values. If categorical columns were included we would then know how many unique options are present. After the more key features were extracted their histograms and box plots were generated as seen in the figures below. These figures were generated to get an idea of the distribution of the data and further understand what the data in these columns looked like. On deeper analysis it appears as if some of the columns have outliers present. Going forwards, these outliers were removed from the dataset. This is because with EEG data the outliers may be looped in with another category of brainwaves. For example, beta waves range from 12-30 Hz and gamma waves from thirty and above. If there was outlier present in the beta wave section it would be looped into the gamma wave classification. I will have to conduct additional research to see how to deal with these outliers.

Figure 4: *Histogram of the most important features*

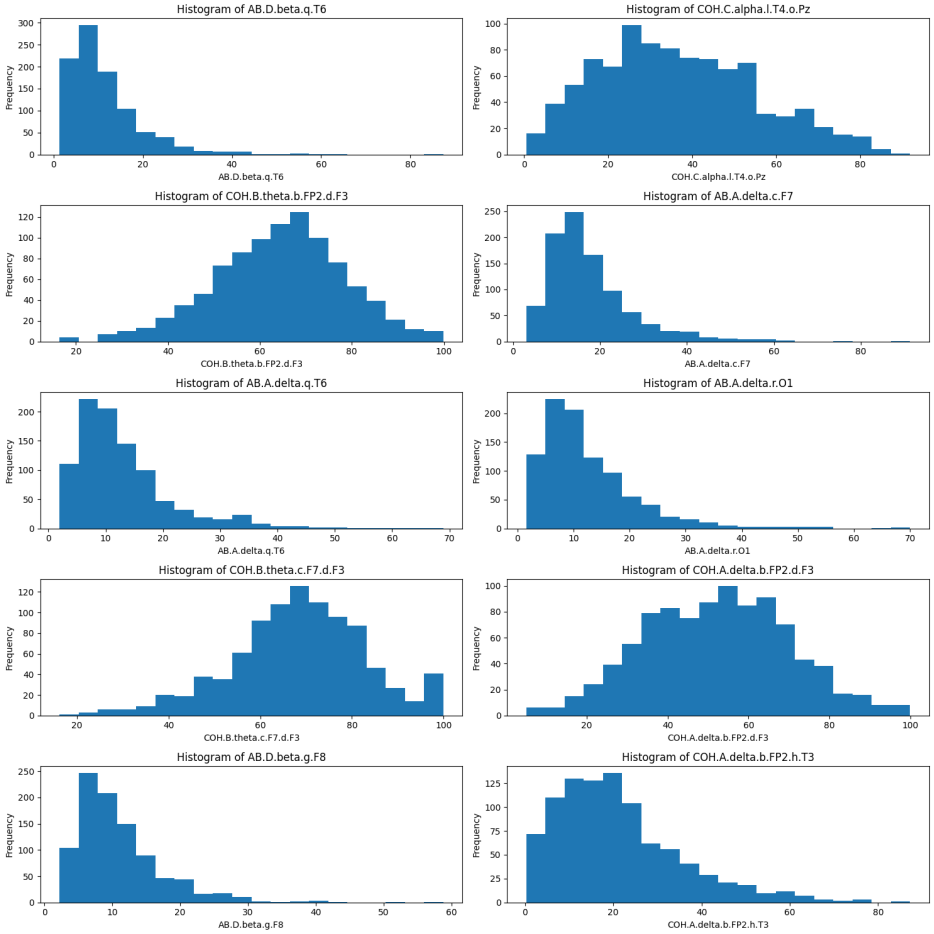
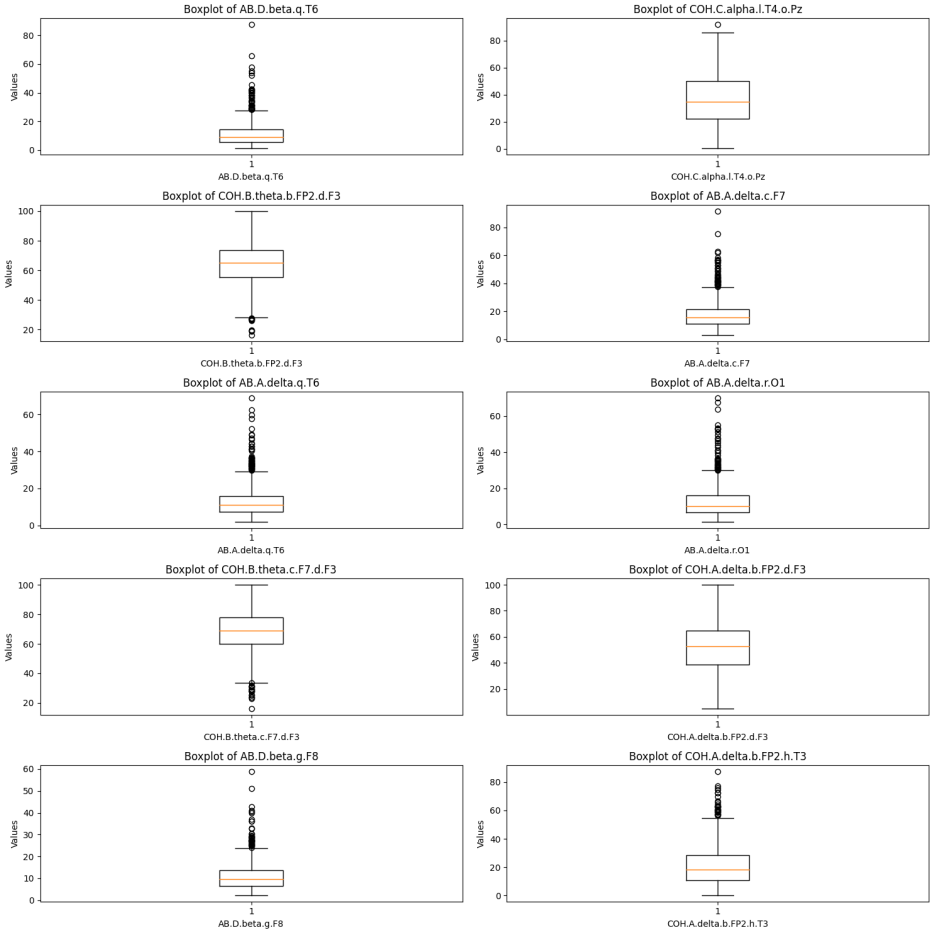
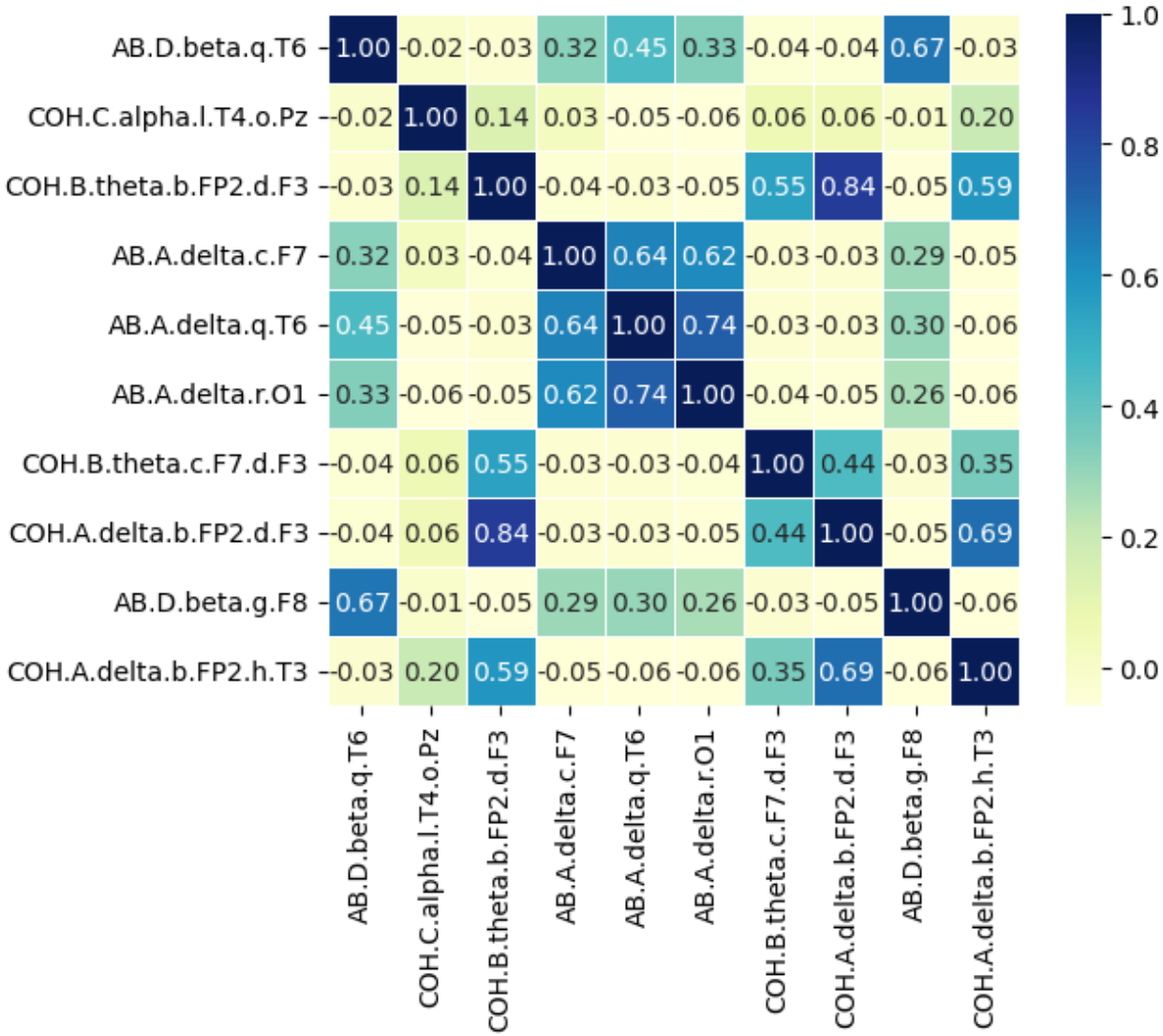


Figure 5: *Box-plots of the most important features*



Next, I conduct multivariate analysis to show the the relationship between the key attributes that we selected from above. Every attribute is plotted against the other and we are able to visualize the relationship thanks to the seaborn correlation matrix. If the coefficient is +1 then the 2 attributes are highly correlated. If the coefficient is -1 then the 2 attributes do not have a relationship. This was one of the cooler visualizations - the correlation heat-map seen below between the important features that were extracted above. By visually representing the correlation coefficients between variables, it helps identify potential patterns, dependencies, and multi-linearity issues, which are essential considerations in feature selection, model building, and data preprocessing tasks. Moreover, the color-coded matrix makes it easy to discern the strength and direction of correlations, with warmer colors indicating stronger positive correlations, cooler colors indicating stronger negative correlations, and neutral colors representing weaker or no correlations. This visual representation facilitates the identification of potentially redundant or highly correlated features, guiding decisions on which features to retain, transform, or discard in subsequent analysis steps.

Figure 6: *Seaborn Correlation Matrix for most important features*



5 Machine Learning Model Selection and Results

Within this section I will be testing different models specifically Support Vector Machines (SVM), Random Forrest Classifier and Convolutional Neural Networks on the cleaned data to see which one works best and what the results are like. To evaluate the model I use a variety of approaches such as classification reports containing overall accuracy, precision, recall. Furthermore I also use confusion matrices and Precision-Recall Curves, when needed, to further evaluate the different machine learning models. It is important to note that in this section the results from the best combination of hyper-parameters was published.

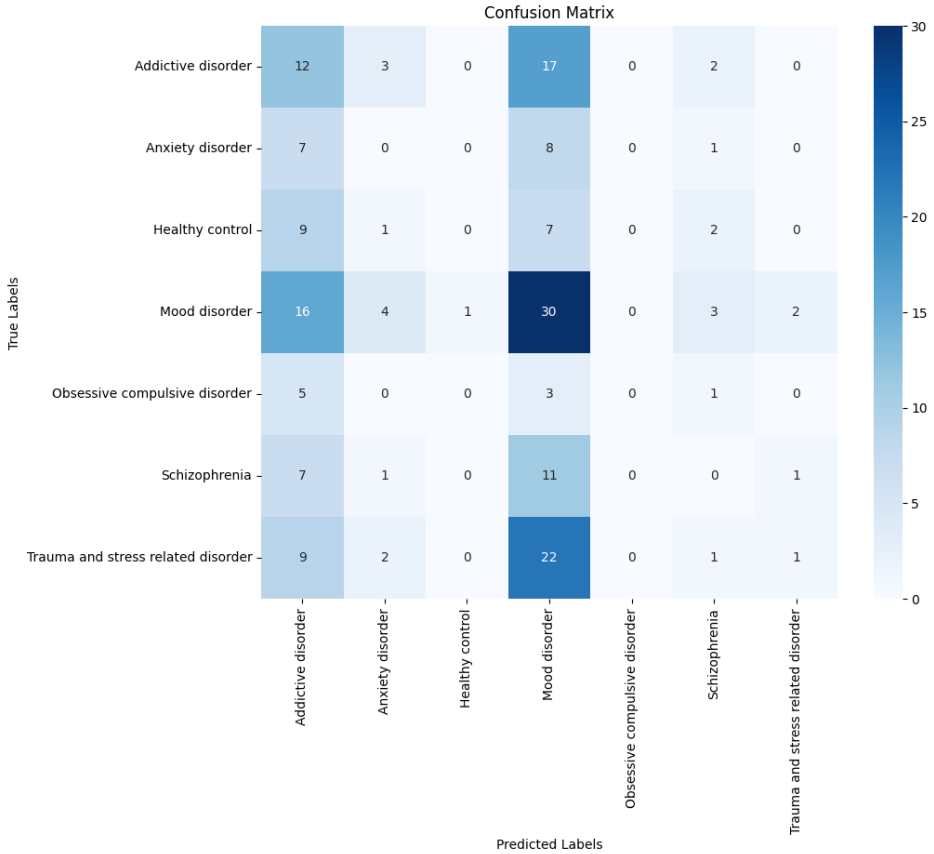
5.1 Support Vector Machine (SVM)

One of the first models that I tried was **Support Vector Machines (SVM)**, I had elected to use this model because it is efficient for non-linear data in finding the best decision boundaries between different classes in a high-dimensional space, making them suitable for this dataset as it is complex and has multiple features. Furthermore, SVMs offer kernels which allow them to manage non-linear data by transforming it into a high-dimensional space where classes can become separable. As seen in the results below, it was not as accurate as predicted (23%). In the report the averages for Precision and Recall are low, indicating that the model did not perform as well as expected. Furthermore, in the confusion matrix of the values used in the test set, the model incorrectly classified most of the disorders as a mood disorder. However, of the disorders that were correctly classified, the model correctly classified Mood disorders as Mood disorders. This can be attributed to the fact that the data is quite imbalanced this is because in the EDA it was clear that the bulk of the data is for Mood Disorders and due to this the model may have been skewed towards classifying more disorders as Mood Disorders.

Figure 7: *Classification Report from best run of SVM algorithm*

```
Best Parameters: {'C': 0.1, 'gamma': 1, 'kernel': 'sigmoid'}
Accuracy: 0.2275132275132275


Classification Report:
                                precision    recall  f1-score   support

             Addictive disorder       0.18      0.35      0.24        34
              Anxiety disorder       0.00      0.00      0.00        16
               Healthy control       0.00      0.00      0.00        19
                 Mood disorder       0.31      0.54      0.39        56
  Obsessive compulsive disorder       0.00      0.00      0.00         9
                  Schizophrenia       0.00      0.00      0.00        20
Trauma and stress related disorder       0.25      0.03      0.05        35

                      accuracy                           0.23       189
                     macro avg       0.11      0.13      0.10       189
                  weighted avg       0.17      0.23      0.17       189
```

Figure 8: *Confusion Matrix showcasing the accuracy of classification.*



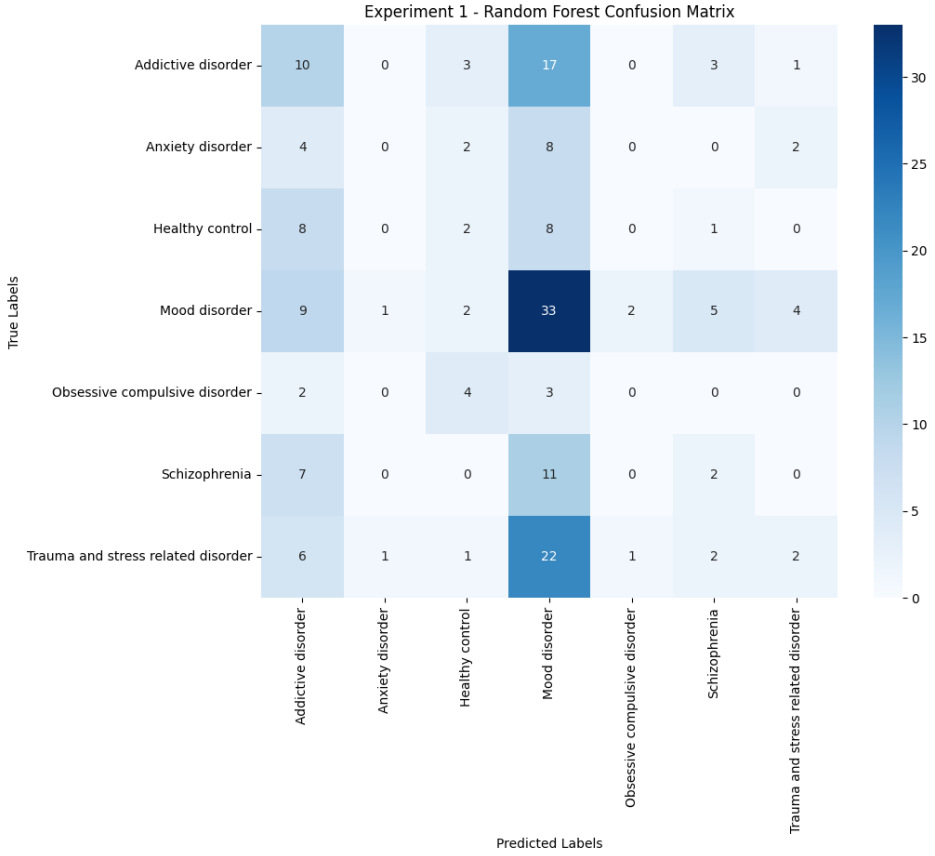
5.2 Random Forrest Classifier

Unfortunately, the SVM model did not perform as well as expected due to the inherent dis-balance that is present in the data. The next model that I decided to use was a Random Forest Classifier. I elected to use this model because based on research it appeared as if Random forests are versatile and often provide reliable performance without much tuning. They work well with both numerical and categorical data and are less prone to over-fitting. When running the model, I ran the experiment with different parameters for n_estimators and max_depth. The first parameter specifies the number of estimators or the number of decision trees that will be built in the random forest, increasing the number of estimators is good, however, having too many will increase computation time and memory requirements and may also reduce the accuracy of the model at times. If plotted it appears as if it will follow a logistic curve and follows the law of diminishing returns. The next parameter is the maximum depth of each decision tree in the random forest. A deeper tree can capture more complex relationships in the data, but it can also lead to over-fitting and more computation time. After running the model and attempting to tune it the greatest accuracy (from experiment 3 in the process_notebook) was ∼30%, making this score slightly higher than the SVM approach and while the averages for the precision and recall scores were higher than that of the SVM, the values are still low indicating that this model also falls short of the goal. In Figure 10 below, in the confusion matrix it can be seen that the Mood Disorders were appropriately classified, this once again is due to the dis-balance present in the data.

Figure 9: *Classification Report from the best run of the Random Forrest Classifier*

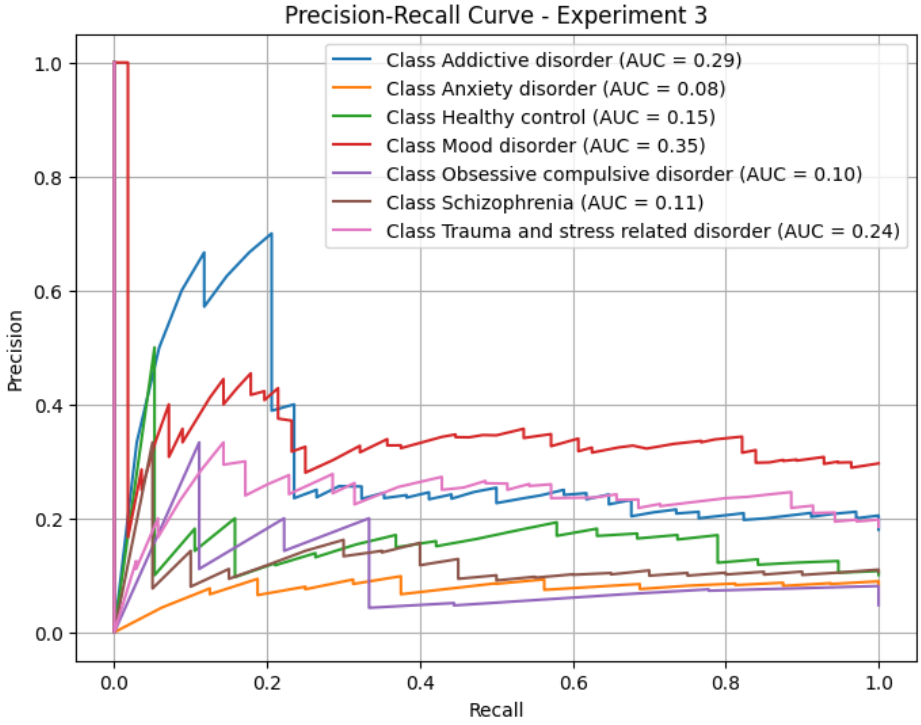
```
Experiment 3 - Random Forest Accuracy: 0.2751322751322751
Experiment 3 - Random Forest Classification Report:
                                precision    recall  f1-score   support

               Addictive disorder       0.30      0.41      0.35        34
                 Anxiety disorder       0.00      0.00      0.00        16
                  Healthy control       0.15      0.11      0.12        19
                    Mood disorder       0.32      0.61      0.42        56
     Obsessive compulsive disorder       0.00      0.00      0.00         9
                      Schizophrenia       0.08      0.05      0.06        20
Trauma and stress related disorder       0.12      0.03      0.05        35

                         accuracy                           0.28       189
                        macro avg       0.14      0.17      0.14       189
                     weighted avg       0.20      0.28      0.22       189
```

Figure 10: *Confusion Matrix showing the accuracy of the Random Forrest Classifier*



Furthermore, I also calculated the Precision-Recall curve. To do this, I had to binarize the data because the precision-recall curve calculation is designed for binary classification. Binarizing the labels involves transforming the multi-class labels into a binary format, where each class is treated as a separate binary classification problem. The AUC in the curve is low for all the different disorders meaning that this model is unable to achieve high precision and recall. Therefore, if not already clear, this model is not effective as a multi-class classification problem. These curves are seen in the figure below.

Figure 11: *Precision-Recall Curve for the Random Forrest Classifier Model*



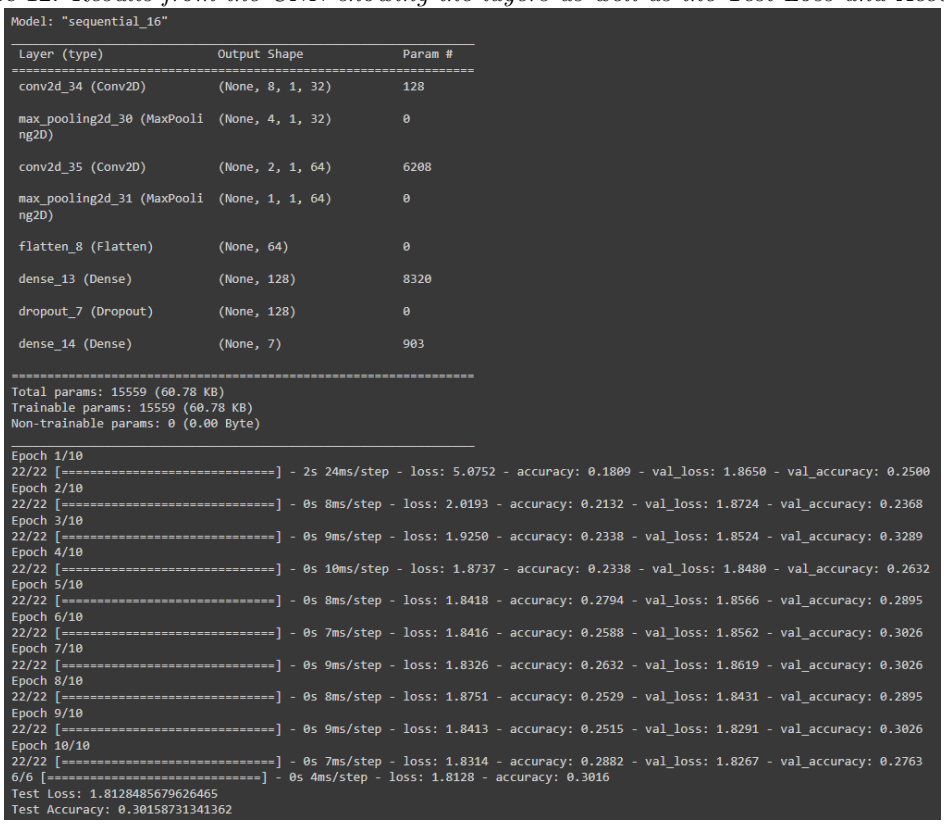
5.3 Convolutional Neural Networks (CNN)

I learned that due to the inherent dis-balance of the data, it is not possible to create a multi-class classification model that can accurately with a 70% score for precision and recall, help classify psychiatric disorders. Next I plan on using Convolutional Neural Networks (CNN), this is because they are well suited to capture spatial and temporal dependencies within data. CNNs are inherently capable of exploiting the spatial structure of multidimensional data, making them ideal for processing EEG signals recorded from different electrodes. Furthermore, CNNs can learn many representations of EEG data. This enables them to apply the model to unseen data, therefore, there is a possibility that the CNN may perform better on the test set. With regards to how much data I will use for this approach? I plan on sticking with the data that was delineated by the feature importance conducted during the EDA. Specifically, I intend to focus on the Coherence metrics, this is because based on a study conducted in 2020, in this study researchers found out that the coherence is reliable for classification because it provides valuable insights into the functional connectivity between brain regions, reflecting the coordination of neural networks involved in various cognitive processes Kamzanova et al. (2020).

The results of the CNN are seen below as before due to the nature of the data being dis-balanced, the CNN did not yield proper results and a multi-class model was not able to be generated with this data.

The accuracy of ∼30% was slightly better but around the same average as yielded by the Random Forest and SVM approaches. However, one approach that I did not consider was to generate a binary classification model. In this binary classification model, given that there is more data for Mood and Addictive disorders, a binary classification model would consist of classifying these 2 vs the rest. The next step would be to further classify these 2 in their independent clusters.

Figure 12: *Results from the CNN showing the layers as well as the Test Loss and Accuracy*



```
Model: "sequential_16"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_34 (Conv2D)          (None, 8, 1, 32)          128

 max_pooling2d_30 (MaxPooli  (None, 4, 1, 32)          0
 ng2D)

 conv2d_35 (Conv2D)          (None, 2, 1, 64)          6208

 max_pooling2d_31 (MaxPooli  (None, 1, 1, 64)          0
 ng2D)

 flatten_8 (Flatten)         (None, 64)                0

 dense_13 (Dense)            (None, 128)               8320

 dropout_7 (Dropout)         (None, 128)               0

 dense_14 (Dense)            (None, 7)                 903

=================================================================
Total params: 15559 (60.78 KB)
Trainable params: 15559 (60.78 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
Epoch 1/10
22/22 [==============================] - 2s 24ms/step - loss: 5.0752 - accuracy: 0.1809 - val_loss: 1.8650 - val_accuracy: 0.2500
Epoch 2/10
22/22 [==============================] - 0s 8ms/step - loss: 2.0193 - accuracy: 0.2132 - val_loss: 1.8724 - val_accuracy: 0.2368
Epoch 3/10
22/22 [==============================] - 0s 9ms/step - loss: 1.9250 - accuracy: 0.2338 - val_loss: 1.8524 - val_accuracy: 0.3289
Epoch 4/10
22/22 [==============================] - 0s 10ms/step - loss: 1.8737 - accuracy: 0.2338 - val_loss: 1.8480 - val_accuracy: 0.2632
Epoch 5/10
22/22 [==============================] - 0s 8ms/step - loss: 1.8418 - accuracy: 0.2794 - val_loss: 1.8566 - val_accuracy: 0.2895
Epoch 6/10
22/22 [==============================] - 0s 7ms/step - loss: 1.8416 - accuracy: 0.2588 - val_loss: 1.8562 - val_accuracy: 0.3026
Epoch 7/10
22/22 [==============================] - 0s 9ms/step - loss: 1.8326 - accuracy: 0.2632 - val_loss: 1.8619 - val_accuracy: 0.3026
Epoch 8/10
22/22 [==============================] - 0s 8ms/step - loss: 1.8751 - accuracy: 0.2529 - val_loss: 1.8431 - val_accuracy: 0.2895
Epoch 9/10
22/22 [==============================] - 0s 9ms/step - loss: 1.8413 - accuracy: 0.2515 - val_loss: 1.8291 - val_accuracy: 0.3026
Epoch 10/10
22/22 [==============================] - 0s 7ms/step - loss: 1.8314 - accuracy: 0.2882 - val_loss: 1.8267 - val_accuracy: 0.2763
6/6 [==============================] - 0s 4ms/step - loss: 1.8128 - accuracy: 0.3016
Test Loss: 1.8128485679626465
Test Accuracy: 0.30158731341362
```

6 Conclusion

Ultimately, it can be concluded that with this dataset the creation of a multi-class classification algorithm is not possible. Based on the results shown above, after testing three different machine learning algorithms it is not possible to generate an accurate model. Going forwards, in order to generate a classification model based on the literature referred to above, as well as my own findings, it is imperative to have a large corpus of diverse data from individuals with different psychiatric disorders. Having a diverse set of data is valuable not only in machine learning, but more so in psychiatry because it ensures that the algorithm and the model can be scaled to clinical applications. Another reason for why having a plethora of data is valuable is because the primary motivation for this study is to aid in psychiatric diagnosis and to reduce the number of misdiagnoses. Ethical considerations must be woven into the creation of the model and the data collection because these are humans whom we are trying to help and slow and careful data collection and data engineering is crucial to fulfill the goal of creating a clinical machine learning model that is used worldwide.

References

Akers, G. (2023). The impact of mental health misdiagnosis. *HillSide*.

Chen, Z. S., Kulkarni, P. P., Galatzer-Levy, I. R., Bigio, B., Nasca, C., and Zhang, Y. (2022). Modern views of machine learning for precision psychiatry. *Patterns*, 3(11):100602.

Dwyer, D. B., Falkai, P., and Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual review of clinical psychology*, 14:91–118.

Kamzanova, A., Matthews, G., and Kustubayeva, A. (2020). Eeg coherence metrics for vigilance: Sensitivity to workload, time-on-task, and individual differences. *Applied psychophysiology and biofeedback*, 45:183–194.

Vermani, M., Marcus, M., and Katzman, M. A. (2011). Rates of detection of mood and anxiety disorders in primary care: a descriptive, cross-sectional study. *The primary care companion for CNS disorders*, 13(2):27211.

(WHO), W. H. O. (2019). Mental health. *WHO*.