Application of Machine Learning Models for Mental Health Classification

Student #23220057
Northeastern University London

***Abstract*— There is immense potential in the ability to apply machine learning models for the prediction of mental health outcomes. This paper, therefore, aims at categorizing major mental health disorders from structured data based on the effectiveness of different ML algorithms. An extensive review of the literature was conducted for assessing current ML models in classification tasks, focusing on mental health prediction. After this the Decision Tree, XGBoost, and SVM classifiers were developed with measures of predictive accuracy. The results indicate that both XGBoost most accurately classified the data in a variety of training scenarios, demonstrating potential for optimizing predictive performance across many diverse datasets and applications. These findings would recommend that the field of mental health could adopt ML tools in general, especially XGBoost, to enable individuals to check their mental health status and appropriate intervention. This would help reduce the workload on mental health professionals and support people who do not have access to therapeutic services.**

Keywords*— *Mental Health Classification, Machine Learning, Decision Tree, XGBoost, Support Vector Machine (SVM), Supervised Learning, Hyperparameter Tuning, Data Privacy, Ethical Implications, Diagnostic Accuracy

I. INTRODUCTION

Mental health significantly impacts emotional, psychological, and social well-being, adversely affecting cognitive functions, emotions, and relationships [1]. Timely assessment is crucial for diagnosing and mitigating these conditions, but often individuals cannot get access to a mental health professional or are unable to recognize that what they are experiencing are symptoms of a mental health condition.

Machine learning (ML), a subfield of artificial intelligence (AI), offers innovative solutions for early identification and intervention in mental health by addressing classification, regression, and clustering problems through data-driven algorithms [2]. By focusing on the correlation between specific symptoms and overall mental health diagnoses, this paper aims to contribute to the development of more efficient, accurate, and accessible mental health diagnostic tools, improving patient care and outcomes.

The effectiveness of ML algorithms, specifically Support Vector Machines (SVM), Decision Trees, and XGBoost, are assessed in classifying mental health conditions from structured symptom data. The dataset includes responses from 120 patients regarding various symptoms used to diagnose mental health conditions, which include Bipolar Type-1 (Mania), Bipolar Type-2 (Depressive), and Major Depressive Disorder, or indicate that the patient does not have a mental health condition (Normal). Subsequent sections include a literature review, data review, model explanations, initial results, model fine-tuning, future applications and development, and ethical implications.

II. LITERATURE REVIEW

A. Overview of Mental Health Issues

Mental health disorders are a significant global health concern affecting millions of people worldwide. According to the World Health Organization (WHO), about 1 in 4 people will be affected by a mental or neurological disorder in their lives, with depression, anxiety, bipolar disorder, and schizophrenia among the most prevalent conditions [3], [4]. The impact of mental health issues extends beyond the individual, affecting families, communities, and even economies through lost productivity and increased healthcare costs [5].

Despite the widespread prevalence of mental health disorders, diagnosing these conditions can be a challenge. Access to mental health services is often limited, especially in rural or underserved areas [6]. Services can be prohibitively expensive and long waitlists delay critical treatment and support. Additionally, the stigma surrounding mental health discourages many from seeking help due to fear of judgment or discrimination [7]. Consequently, there is a growing need for innovative solutions which can enhance early detection, improve diagnostic accuracy, and provide more accessible and cost-effective mental health assessments.

B. Introduction to Machine Learning

ML involves the development of algorithms that can learn from and make predictions based on data. In recent years, ML has shown great promise in various fields, including healthcare [8]. By analyzing large datasets, MLalgorithms can identify patterns and make predictions with a level of accuracy that often surpasses traditional methods.

In the context of mental health, ML models can analyze diverse data sources, such as electronic health records (EHRs), genetic information, and even social media activity, to predict the likelihood of mental health issues and suggest appropriate interventions [9]. This approach can enhance diagnostic accuracy and enable early detection and personalized treatment plans [10].

C. Existing Applications of Machine Learning in Mental Health

ML has been increasingly applied to diagnose and predict mental health conditions. For instance, the utilization of EHRs to predict the onset of depression with ML algorithms demonstrated that these models could achieve high accuracy, surpassing traditional diagnostic methods [11].

Another study employed natural language processing (NLP) techniques to analyze unstructured clinical notes and identify patients at risk of suicide. It incorporated demographic data, medical history, and textual data to achieve significant predictive performance [12].

Additionally, the analysis of social media posts to detect signs of depression successfully identified markers of depression in digital activity [13]. These studies collectively illustrate how ML can enhance diagnostic accuracy, enable early intervention, and leverage diverse data sources to improve mental health outcomes.

III. Methodology

A. Mental Disorder Classification Dataset

The dataset comprises 120 samples, 17 features, and 4 classes. The samples are responses from psychology patients rating their experience with the 17 features, which are symptoms used to diagnose the 4 classes, which are mental health conditions. The symptoms include Sadness, Exhaustion, Euphoria, Sleep disorder, Mood swings, Suicidal thoughts, Anorexia, Anxiety, Try-explaining, Nervous breakdown, Ignore & Move-on, Admitting mistakes, Overthinking, Aggressive response, Optimism, Sexual activity, and Concentration. The mental health conditions are Bipolar Type-1 (Mania), Bipolar Type-2 (Depressive), Major Depressive Disorder, and Normal individuals. The data is almost perfectly evenly split between the 4 classes, with each class comprising ~25% of total instances. The data is presented in a Comma Separated Value (CSV) format. The data was sourced from the public domain and is available on Kaggle [14].

B. Software and Libraries

This paper utilizes Python 3 for all aspects, including data preprocessing, model development, and evaluation. Key libraries include pandas for data manipulation, NumPy for numerical operations, and scikit-learn for model development and evaluation. Specific tools from scikit-learn include train_test_split, LabelEncoder, DecisionTreeClassifier, SVC, XGBoost's XGBClassifier, GridSearchCV, and classification_report. Matplotlib, seaborn, and LaTex are used for data visualization. Jupyter Notebooks are used for the integrated development environment (IDE).

C. Model Selection

Support Vector Machines (SVM), Decision Trees, and XGBoost were selected for their proven effectiveness in similar mental health diagnostic studies [2], [15], [16], [17], [18]. SVMs are ideal for binary classification tasks and handle high-dimensional data well, making them suitable for datasets with complex, non-linear relationships. Decision Trees are chosen for their interpretability and ability to manage categorical data, providing clear decision rules crucial in clinical settings. XGBoost, known for its high performance and robustness, efficiently handles complex feature interactions and prevents overfitting through advanced regularization techniques.

To identify the most effective model for diagnosing mental health conditions, the performance of Support Vector Machines (SVM), Decision Trees, and XGBoost were compared. The training set and test set were fixed as 75:25 respectively. Each model was trained and evaluated on the dataset. Classification reports were generated for all three models to assess their accuracy, precision, recall, and F1 scores, providing a comprehensive evaluation of each model's ability to correctly classify mental health conditions. The Decision Tree and XGBoost model performed equally with an 75:25 data split, so further split ratios were tested. XGBoost consistently performed better across all ratios so was selected for further optimization and deployment.

D. Model Tuning and Deployment

After selection, the XGBoost model underwent hyperparameter tuning to optimize its performance. A grid search with 5-fold cross-validation was performed to evaluate hyperparameters. In this process, the dataset was divided into five parts, the model was trained on four parts, and validated on one part. The combination of hyperparameters yielding the highest cross-validation score was selected for the final model. The model was then retrained on the general dataset and deployed to the data. The resulting performance was measured to achieve the aspect of consistency and reliability.

IV. Results

A. Initial Model Comparison

The performance of measures of classifiers measures the decision-making capability of the classifier. The measures used to determine the performance are accuracy, precision, recall, F-score.

Accuracy gives an overall effectiveness of a classifier. The accuracy scores obtained for the classifiers built are given in Table 1. Decision Tree and XGBoost performed the highest with a score of 0.73, in that they correctly classified 73% of instances.

	SVC	**Decision Tree**	**XGBoost**
Accuracy	0.67	0.73	0.73

Table 1: Accuracy

Precision is a measure of the class agreement of the data labels with the positive labels given by the classifier. Table 2 shows the values for precision score for each classifier and label. Decision Tree and XGBoost scored highest with a 0.76 macro average for both, indicating that 76% of positive identifications were correct.

	SVC	Decision Tree	XGBoost
Bipolar Type-1	0.45	0.50	0.50
Bipolar Type-2	0.50	1.00	1.00
Depression	1.00	0.78	0.78
Normal	1.00	0.75	0.75

Table 2: Precision

Recall represents the classifier's effectiveness to identify class labels. The recall scores for 3 class labels and the classifiers are shown in Table 3. Decision Tree and XGBoost again scored highest at 0.75 macro average, indicating that the model correctly identified the instances of the target class 75% of the time.

	SVC	Decision Tree	XGBoost
Bipolar Type-1	1.00	0.80	0.80
Bipolar Type-2	0.67	0.83	0.83
Depression	0.56	0.78	0.78
Normal	0.60	0.60	0.60

Table 3: Recall

Lastly, F- Score gives the relationship between positive labels and those given by the classifier. It is computed by taking the harmonic mean of precision and recall for all the 3 labels across all the classifiers. The F scores for the class labels are shown in Table 4. Again. the classifiers Decision Tree and XGBoost scored highest at 0.75 macro average

	SVC	Decision Tree	XGBoost
Bipolar Type-1	0.62	0.62	0.62
Bipolar Type-2	0.57	0.91	0.91
Depression	0.71	0.78	0.78
Normal	0.75	0.67	0.67

Table 4: F1-Score

Since both models performed equally for all metrics, the ratio by which the train and test data was split was modified to assess performance of both models with more and less training and testing data. A split of 80:20 yielded the same results, while splits of 85:15 and 70:30 showed significant differences, with XGBoost outperforming Decision Tree by an average of 0.1 and 0.05 for all metrics, respectively.

	70:30	85:15
Decision Tree	0.79	0.66
XGBoost	0.84	0.75
Difference	0.05	0.10

Table 5: Average of all classification report values

XGBoost was therefore chosen for further tuning and deployment as it performed better on multiple data splits, has generally better performance due to advanced techniques, and is not as prone to overfitting as Decision Tree. This is especially important as the model utilizes medical data and provides a diagnosis, where accuracy and reliability are crucial for patient outcomes.

B. XGBoost Tuning and Deployment

To tune the XGBoost model, a grid search with 5-fold cross-validation was performed to evaluate hyperparameters.
The hyperparameters evaluated were n_estimators, max_depth, learning_rate, subsample, colsample_bytree, gamma, reg_alpha, reg_lambda, booster, tree_method, n_jobs, and random_state. The highest cross-validation score achieved was 0.91, but upon application of the optimized hyperparameters, the model did not improve. There was an attempt to test more hyperparameters but there was not sufficient computing power. Manual adjustment of certain hyperparameters achieved an improved model, indicating that further analysis of hyperparameters may achieve a better performing model.

The normalized confusion matrix in Fig. 1 displays the classification performance of a model across the four classes. The values in the matrix represent the proportion of correct and incorrect predictions normalized by the actual class counts. For Bipolar Type-1, 80% of cases were correctly classified, while 20% were misclassified as Normal. In the case of Bipolar Type-2, 83% were correctly identified, with 17% cases being classified as Bipolar Type-1. For Depression, 78% of cases were correctly classified, with 11% of cases misclassified as Bipolar Type-1 and another 11% as Normal. As for the Normal category, 60% of cases were correctly classified, with 20% being classified as Bipolar Type-1 and another 20% as Depression. Overall, the model shows reasonable performance in classifying the given categories, though there are some notable misclassifications, particularly between Bipolar Type-2 and Bipolar Type-1, and between Normal and the other conditions.

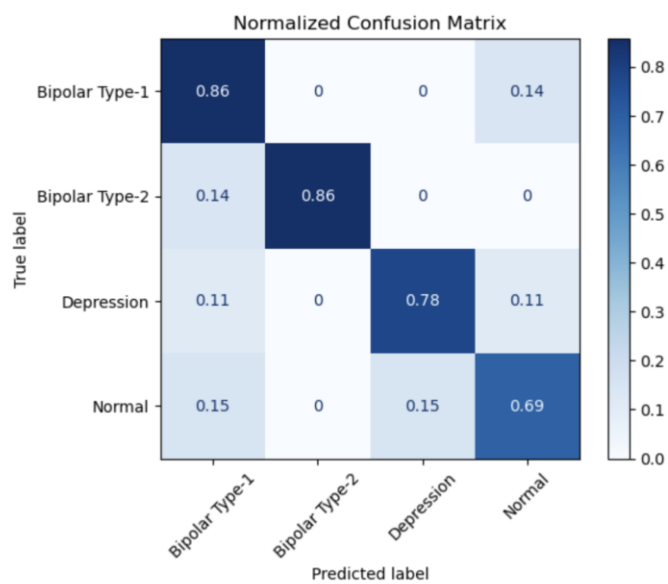

Fig. 1

C. Drawbacks and Further Analysis

While this model performs well, significant improvement is needed before considering its use in a clinical setting. Firstly, a much larger dataset is required to give ample training data to the model. Currently there are only 120 samples, which are further split into the training and test sets. This limited sample size restricts the model's ability to accurately predict diagnoses.

Additionally, the grid search did not yield any significant improvement to the model. More hyperparameters were attempted to be tested, but there was not sufficient computing power. Running the model through a more powerful machine may yield more options. The model can also be examined for potential overfitting or underfitting by comparing training and validation scores and using techniques like regularization or model simplification if needed.

With medical diagnosis, models must perform well to not misdiagnose patients, since this could lead to patients not receiving proper treatment or being treated for a condition they don't have. XGBoost has been shown to be viable option for this application among models that utilize supervised learning, and with sufficient data, computing power, and further analysis could be an efficient tool for diagnosing mental health conditions. However, it would be beneficial to compare performance to models that utilize unsupervised learning or deep learning methods such as recursive neural networks. This comparison could provide a more comprehensive evaluation of model effectiveness and potentially highlight different strengths and weaknesses in diagnostic capabilities.

V. Ethical Considerations

The development of ML algorithms for diagnosing mental health conditions offers significant potential for improving patient care and outcomes. However, ethical considerations must be addressed to ensure responsible use. Data privacy is paramount; although the dataset that was utilized is anonymized, there is no information as to the original anonymization process, which raises concerns about the adequacy and effectiveness of the methods used.

In a healthcare application such as this, mitigating risks like misdiagnosis is essential. Continuous validation of algorithm accuracy is necessary in ensuring that these tools supplement professional diagnoses rather than replace them. The impact on stakeholders includes providing patients with easier access to diagnostic services, which may lead to reliance on the tool and avoidance of professional help. For healthcare providers, these tools can enhance diagnostic accuracy but must support, not substitute, professional judgment.

Accessibility is a concern, as this technology requires devices like smartphones or computers to be utilized.

Compliance with regulations such as HIPAA and GDPR is vital. This includes regular audits, data protection impact assessments, and obtaining informed consent wheremaci applicable. Ensuring fairness and reducing bias in algorithms is crucial. This also involves conducting regular audits, as well as using diverse training data and maintaining transparency in development.

Addressing these ethical considerations is critical for the successful integration of ML in mental health diagnostics. This assures that the technology provides an accurate and reliable assessment and upholds principles of fairness, privacy, and transparency that foster trust and acceptance among users and practitioners.

References

[1] World Health Organization, "Mental health: strengthening our response," Who.int, Jun. 17, 2022. https://www.who.int/en/news-room/fact-sheets/detail/mental-health-strengthening-our-response

[2] A. Thieme, D. Belgrave, and G. Doherty, "Machine Learning in Mental Health," ACM Transactions on Computer-Human Interaction, vol. 27, no. 5, pp. 1–53, Oct. 2020, doi: https://doi.org/10.1145/3398069.

[3] J. Sayers, "The world health report 2001 — Mental health: new understanding, new hope," Bulletin of the World Health Organization, vol. 79, no. 11, p. 1085, 2001, Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2566704/

[4] *GBD 2019 Mental Disorders Collaborators, "Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019," The Lancet Psychiatry, vol. 9, no. 2, pp. 137–150, Jan. 2022, doi: https://doi.org/10.1016/s2215-0366(21)00395-3.*

[5] R. C. Kessler, P. Berglund, O. Demler, R. Jin, K. R. Merikangas, and E. E. Walters, "Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication," Archives of General Psychiatry, vol. 62, no. 6, pp. 593–602, Jun. 2005, doi: https://doi.org/10.1001/archpsyc.62.6.593.

[6] S. Chaturvedi, Ed., Mental Health and Illness in the Rural World. Singapore: Springer, 2020. doi: 10.1007/978-981-10-2345-3

[7] Arboleda-FlórezJ. and N. Sartorius, Understanding the stigma of mental illness : theory and interventions. Chichester, England: John Wiley & Sons, 2008.

[8] A. Esteva et al., "A Guide to Deep Learning in Healthcare," Nature Medicine, vol. 25, no. 1, pp. 24–29, Jan. 2019, doi: https://doi.org/10.1038/s41591-018-0316-z.

[9] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: a scoping review of methods and applications," Psychological Medicine, vol. 49, no. 09, pp. 1426–1448, Feb. 2019, doi: https://doi.org/10.1017/s0033291719000151.

[10] T. Davenport and R. Kalakota, "The Potential for Artificial Intelligence in Healthcare," Future Healthcare Journal, vol. 6, no. 2, pp. 94–98, Jun. 2019, doi: https://doi.org/10.7861/futurehosp.6-2-94.

[11] D. Nickson, C. Meyer, L. Walasek, and C. Toro, "Prediction and diagnosis of depression using machine learning with electronic health records data: a systematic review," BMC Medical Informatics & Decision Making, vol. 23, no. 1, pp. 1–38, Nov. 2023, doi: https://doi.org/10.1186/s12911-023-02341-x.

[12] M. Adamou, G. Antoniou, E. Greasidou, V. Lagani, P. Charonyktakis, and I. Tsamardinos, "Mining Free-Text Medical Notes for Suicide Risk Assessment," Proceedings of the 10th Hellenic Conference on Artificial Intelligence, Jul. 2018, doi: https://doi.org/10.1145/3200947.3201020.

[13] A. G. Reece, A. J. Reagan, K. L. M. Lix, P. S. Dodds, C. M. Danforth, and E. J. Langer, "Forecasting the onset and course of mental illness with Twitter data," Scientific Reports, vol. 7, no. 1, Oct. 2017, doi: https://doi.org/10.1038/s41598-017-12961-9.

[14] "Mental Disorder Classification," www.kaggle.com. https://www.kaggle.com/datasets/cid007/mental-disorder-classification

[15] M. Srividya, S. Mohanavalli, and N. Bhalaji, "Behavioral Modeling for Mental Health using Machine Learning Algorithms," Journal of

Medical Systems, vol. 42, no. 5, Apr. 2018, doi: https://doi.org/10.1007/s10916-018-0934-5.

[16] H. Yang and P. A. Bath, “Automatic Prediction of Depression in Older Age,” Proceedings of the third International Conference on Medical and Health Informatics 2019 - ICMHI 2019, 2019, doi: https://doi.org/10.1145/3340037.3340042.

[17] A. Thieme, D. Belgrave, and G. Doherty, “Machine Learning in Mental Health,” ACM Transactions on Computer-Human Interaction, vol. 27, no. 5, pp. 1–53, Oct. 2020, doi: https://doi.org/10.1145/3398069.

[18] P. P. Shinde and S. Shah, “A Review of Machine Learning and Deep Learning Applications,” IEEE Xplore, Aug. 01, 2018. https://ieeexplore.ieee.org/document/8697857