Predicting Mental Health Disorders Using Client Level Reporting Data and Machine Learning Models

Daniel Harvey
Columbia University
dyh2111@columbia.edu

Dr. Michelle F. Levine
Columbia University
mfl41@columbia.edu

Abstract

This study explores the use of machine learning to predict mental health disorders using the 2022 U.S. Department of Health and Human Services' CBHSQ Mental Health Client-Level Data (MH-CLD). The dataset (N = 370,515) was filtered to focus on a representative adult population aged 18 and older who have completed high school and meet specific societal criteria. Key demographic and socioeconomic features including age, gender, race, marital status, education, and employment, were utilized to predict one of five mental health disorders including anxiety, bipolar disorder, and schizophrenia.

Multiple machine learning models were tested and evaluated, with recall chosen as the primary performance metric to assess the models' effectiveness in identifying cases. To address class imbalance, the dataset was upsampled using SMOTE)=. Among the tested models, a seven-layer feed-forward neural network (N_parameters = 100,000) achieved an average recall rate positive classification rate of 95.8%, demonstrating the potential of machine learning in early detection and prediction of mental health disorders in the studied population.

1 Introduction

Mental health has become an increasingly prominent topic in everyday life, gaining attention in schools, workplaces, and the media. Discussions about mental health are now common in college orientations and on social platforms. However, many mental health disorders often go undetected or unrecognized until they reach a critical stage, thus having a tool for early detection would be crucial for improving long term quality of life. This is particularly important given the various levels of impairment and debilitation mental illnesses can cause.

It is estimated that over one in five U.S. adults live with a mental illness, accounting for 59.3 million individuals in 2022, or 23.1% of the adult population (NIMH, 2022).

Mental illnesses encompass a wide range of conditions that require clinical diagnosis, contingent on the patient recognizing the need for help and seeking a provider. In the United States, diagnoses are typically guided by two key classification systems: the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the International Classification of Diseases (ICD). These systems categorize mental, behavioral, and emotional disorders and cover a broad spectrum of conditions, including anxiety, depression, and schizophrenia (Clark, L. A., 2017).

Despite their utility, accurately diagnosing mental illnesses remains challenging due to issues such as overlapping symptoms among disorders (comorbidity), the arbitrary thresholds defining specific conditions, and the complexity of their multifactorial etiology. While the DSM and ICD provide essential frameworks, their limitations underscore the need for complementary tools to address the high prevalence and diverse presentations of mental illnesses.

2 Research Question

This study leverages the U.S. Department of Health and Human Services' Mental Health Client-Level Data (MH-CLD) to predict the likelihood of developing specific mental health disorders using machine learning models. At the time of this study, the most recent publicly available dataset available through the SAMHSA website, the 2022 MH-CLD dataset containing records (N=6,957,919) of clinically diagnosed individuals, the study seeks to explore the predictive power of demographic features such as age, gender, race, location, education, marital status, and employment in mental health diagnosis. The goal is to determine whether disorders such as trauma and stressor related disorders, anxiety, bipolar, depressive, and schizophrenia/psychotic can be accurately classified, ultimately leading to the notion of correlation to a diagnosis. Machine learning models trained on this dataset provide an opportunity to uncover complex patterns and predictors that could aid in early detection or at least grounds for a consult, especially in low threshold patients who do not yet realize their disorder, ultimately contributing to improved quality of life and reduced long-term care requirements for affected individuals.

Additionally, with the rise of social media—particularly driven by lifestyle, travel, and business influencers—and given that this data was collected toward the end of the COVID-19 pandemic, we hypothesize that self-esteem and self-worth-related mental health issues, such as depression, would be most prevalent in the general public, particularly among unemployed individuals.

3 Related Work

Despite the large sample size and high-quality, provider-sourced data in the SAMSHA MH-CLD, this dataset remains underutilized in peer-reviewed scientific research. Where some studies have leveraged it, our project's scope and target populations differ from the present work.

In one study, a multi-label neural network was applied to predict mental health outcomes in three categories of young adults aged 15 to 24 focusing on using age, education, race, gender, marital status, and their employment as predictive features for predicting Anxiety and Depression (Verma and Supekar, 2024). Their model performed well, achieving an average accuracy of 93%. However, we noted that accuracy was the only evaluation metric used. In the unfiltered dataset, depressive disorders accounted for 26.6% of cases, while anxiety made up 19.8%. Given this class imbalance, achieving a high accuracy score could be misleading. For example, in the case of anxiety, a model that classified every case as negative would still achieve 80.2% accuracy. This reliance on accuracy alone risks obscuring the model's performance in identifying the true target disorders.

Another study explored the use of neural networks and logistic regression to examine the co-occurrence of substance abuse with anxiety and depressive disorders in adults (Ware et al., 2024). This study found that approximately 30% of the dataset's population exhibited co-occurrence, with key predictive factors including the region where treatment was received, age, education, gender, race, ethnicity, and the presence of anxiety and depressive disorders. While this study utilized the same dataset, it focused on the interplay between mental health and substance use—a population explicitly excluded from the present study.

A study examining the prevalence of mental health disorders in older adults 50+ found this age group had higher odds of developing depressive disorders compared to younger age groups (Choi 2022). Utilizing logistic regression, researchers identified gender, race/ethnicity, census region, and alcohol/substance use disorder as significant predictive features. Additionally, their study found that this older population specifically aged 50–64 and 65+ had higher odds of depressive disorders in outpatient-only settings and were more likely to be diagnosed with schizophrenia or other psychotic disorders across outpatient-only, combined outpatient and inpatient, and inpatient-only service settings.

A study published in 2023 investigated the relationship between trauma and stressor related disorders and substance abuse vs gender (Ware 2023). By utilizing a logistic regression model together with the 2013 - 2019 MH-CLD datasets, they found that men were more likely to have substance use together with serious mental illnesses.

Lastly, a study examining a diverse array of studies utilizing several data sources, including social media interactions, genetic profiles, electroencephalogram (EEG) data, and electronic health records, discussed the growing body of work utilizing machine learning in the mental health realm (Su, C et al 2020). With 57 out of 2261 studies meeting the inclusion criteria, the authors discussed the promising work being produced but also the pitfalls of these complex, more difficult to explain models.

In this study, we focus on a refined population group: adults aged 18 and older who meet stricter criteria emphasizing societal factors such as education level, stable housing, and the absence of substance abuse issues. This approach aims to model the general workforce-eligible population. Additionally, by implementing neural networks with class-level granular metrics and concentrating on a distinct population segment not addressed in previous MH-CLD-based research, this study provides a novel contribution to predictive mental health analysis.

4 Dataset

The dataset used in this study originates from the U.S. Department of Health and Human Services' Substance Abuse and Mental Health Services Administration (SAMHSA), which oversees public behavioral health. Public and private mental health care providers receiving public funding report client level data on individuals receiving mental health treatment services during a state-defined 12-month reporting period. This data is provided processed, cleaned, and anonymized by SAMHSA for public release, such that the dataset includes only non-personally identifiable information, such as up to three clinically diagnosed mental health conditions, patient demographics, and data for new, current, and discharged patients.

It encompasses data from all reporting states as well as the Commonwealth of the Northern Mariana Islands, the Republic of Palau, Puerto Rico, and the District of Columbia. For 2022, locales such as American Samoa, the Federated States of Micronesia, Guam, Maine, the Marshall Islands, and the U.S. Virgin Islands did not provide sufficient data to be included in this edition, but have contributed to past editions.

Each record represents an individual under care, with 40 features encoded numerically using a feature map. To focus on the risks of mental health disorders within the typical adult, work-able population, an a priori inclusion criteria was developed to include individuals who are 18 years or older, have completed high school or equivalent, reside in private housing, and do not have a reported substance abuse problem. The filtering criteria were defined in Table 1.

5 Tech Stack

After filtering, the final dataset comprised 370,817 samples spanning 10 features. The large volume of data was processed using Google Cloud's BigQuery, a SQL-based data warehousing platform, while Python 3.10 and Colab Enterprise using either an NVIDIA T4 or A100 GPU were employed for data exploration, modeling, and prediction.

Data analysis revealed significant class imbalances among the 12 mental health disorders. The most prevalent conditions were depressive disorders (33.4%), bipolar disorder (18.1%), and schizophrenia/psychotic disorders (15.1%). In contrast, the least represented disorders—conduct disorder, delirium/dementia, and oppositional defiant disorder—each accounted for less than 0.1% of cases. Due to these extreme imbalances, the original plan to predict the most likely mental disorder among all 12 classes was reconsidered. This study instead focuses on the five most represented disorders: depressive, bipolar, schizophrenia/psychotic, anxiety, and trauma- and stressor-related disorders.

6 Metrics

Given the medical nature of the research question and the recognition that accuracy is not an appropriate metric due to the imbalanced data, the priority was to minimize false negatives, as missed diagnoses would have greater implications as we posed this study as an early detection tool. Consequently, recall was chosen as the primary evaluation metric. The neural network demonstrated the greatest potential for improving recall and was selected for further optimization.

7 Baseline Model

To ensure compatibility with various machine learning models during selection, the features were transformed into one-hot encoded vectors. The dataset was split into development, test, and validation sets using a 60/30/10 ratio, with stratification to preserve even class distribution across the splits. For initial model exploration, a smaller stratified subset (N = 74,103) was created, maintaining the original feature distribution. This subset was further divided into binary classification tasks for each mental illness, addressing class imbalance issues and enabling more reliable predictions during the initial model selection phase.

The depressive disorders subset, having the highest number of cases (N = 74,910), was selected to evaluate potential models. A logistic regression model with L1 regularization was implemented as a baseline for each of the five disorders, providing initial performance benchmarks. Additionally, several unoptimized models, including decision trees, support vector machines (SVMs) implemented with Scikit-learn, and a fully connected neural network built with TensorFlow/Keras were all tested for model consideration.

8 Final Model

The neural network offers several advantages. According to the Universal Approximation Theorem (Hornik et al 1989), neural networks have the ability to model any distribution or mathematical function. Additionally, when new data becomes available, they can continue training or be retrained. As historical datasets from 2012 were available, plans to incorporate this data to assess whether it could improve predictive performance, as neural networks generally benefit from more data.

To avoid the inconvenience of building, tuning, and managing five separate models for each diagnosis, five one-hot encoded features were added to represent the type of diagnosis. This allows a single model to classify and switch modes depending on the diagnosis being considered.

Model	**Recall**	**F1-Score**	**Accuracy**
Logistic Regression	0.60	0.49	0.58
LR - Elastic	0.60	0.49	0.58
LR Newton-Colesky	0.60	0.49	0.58
Decision Tree	0.57	0.47	0.57
Random Forest	0.55	0.47	0.58
Neural Network	0.63	0.49	0.55

Table 2: Exploratory Model Results for Depressive Disorder

The model consists of a 7-layer feed-forward fully connected neural network that employs dropout regularization. The model was tuned across several hyperparameters to find the optimal configuration based on the validation set. The model's output is binary, utilizing a learning rate of 0.012, binary cross-entropy loss, and a final binary sigmoid output.

The final model was trained on the entire dataset, but its performance remained suboptimal due to the class imbalance. To address this, a synthetic resampling technique called Borderline-SMOTE was applied only to the training data. SMOTE (Synthetic Minority Over-sampling Technique) is a method that generates synthetic samples for the minority class to balance the dataset by modeling its distribution to match that of the majority class. With SMOTE, the new dataset (N=1816418) was created and used for training. Imbalance techniques should be applied only to the training data when training a model, while the original, unaltered data should be used for validation and testing to accurately assess the model's generalization to future data.
The model was trained for 25 epochs with a batch size of 128, optimizing recall on the validation set.

Feature	Description
Age	Limited to individuals 18+ to exclude developmental differences in younger patients (0–11 years), which may affect symptom reporting and self-expression.
Education	Included only high school graduates or above to reflect a typical work-eligible population, excluding individuals who may have faced educational barriers.
Race	Records missing race information were excluded, leaving approximately 6 million records. Including race enhances predictive accuracy for population trends.
Gender	Invalid or missing gender data were excluded, retaining about 6.9 million records without compromising data integrity.
State	Non-U.S. jurisdictions (853 records) were excluded to focus on the general U.S. population.
Marital Status	While marital status data were missing for 40% of records, the remaining data offer valuable insights into demographic trends.
Employment Status	Excluded records with 'Unknown' employment status (61% of data) to focus on employment's correlation with mental health.
Housing	Retained only records for individuals living in private residences, aligning with the study's focus on typical societal contributors.
Number of Diagnosed Disorders	Retained all records as these provide direct insights into mental health burdens.
Substance Abuse	Excluded individuals with reported substance abuse problems to reduce confounding factors.

Table 1: Filtering Criteria of Dataset

9 Results

The model demonstrated an average recall score of 0.75, F1-score of 0.42 across all disorders. These findings suggest that neural networks hold promise for predicting mental health conditions using broad feature sets.

10 Future Work

Given the promising initial results, future work could expand the scope and utility of the model in several ways. Incorporating top-3 predictions based on the highest probabilities could align more closely with the dataset, where up to three conditions are often reported. Feature expansion, such as integrating additional data sources like electronic health records, vital signs, and laboratory results, could further enhance the model's predictive accuracy by capturing clinical and physiological markers. Additionally, while socioeconomic and demographic data offer valuable insights, their predictive power may be limited without more granular details. Collecting patient-level data, including lifestyle habits, social factors, and psychological assessments, could significantly improve the model's ability to identify and predict mental health conditions

11 Limitations

This study relied exclusively on data provided by the Substance Abuse and Mental Health Services Administration (SAMHSA). While this dataset is extensive, it represents only individuals who received funding for mental health services through this organization. As such, the data may not fully reflect the broader population, particularly individuals receiving care from private providers who may have different socioeconomic and demographic profiles. This discrepancy could influence the distribution of mental health disorders and limit the model's generalizability.

Additionally, the dataset inherently excludes individuals who are unaware of their mental health conditions or those who lack access to, or choose not to seek, professional help. This omission highlights the potential for bias in the results, as the data does not capture the full spectrum of mental health conditions in the general population.

Finally, it is important to acknowledge that the machine learning tools used in this study are intended solely for educational and research purposes. They are exploratory in nature and not designed for diagnostic use. Accurate diagnosis of mental health conditions should always be performed by licensed professionals.

Ethics Statement

The dataset used in this study was provided and approved by the Substance Abuse and Mental Health Services Administration (SAMHSA). It was fully anonymized and prepared in accordance with guidelines to ensure the protection of personally identifiable information (PII), making it suitable for public consumption.

This paper was not reviewed or approved by the Columbia ethics board. The analysis and interpretations presented are solely the responsibility of the authors and were conducted independently within the scope of public data usage.

Acknowledgments

The author extends their gratitude to Professor Levine for her invaluable guidance throughout the course Empirical Methods of Data Science. Special thanks are also due to TA Aimee Oh for her support and to fellow classmates for their collaboration, peer-review, and encouragement.

References

Clark, L. A., Cuthbert, B., Lewis-Fernández, R., Narrow, W. E., & Reed, G. M. (2017). Three Approaches to Understanding and Classifying Mental Disorder: ICD-11, DSM-5, and the National Institute of Mental Health's Research Domain Criteria (RDoC). *Psychological Science in the Public Interest*, 18(2), 72-145. https://doi.org/10.1177/1529100617727266

Choi, N. G., DiNitto, D. M., & Marti, C. N. (2022). Public mental health service use among U.S. adults age 50+ compared to younger age groups. *Social Work in Health Care*, 61(9–10), 499–515. https://doi.org/10.1080/00981389.2022.2154886

Edward, V., & Moreira, V. (2023). Mental Health in Racial Minorities. *Digital Commons Kennesaw*. Available at: https://digitalcommons.kennesaw.edu

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359-366. https://doi.org/10.1016/0893-6080(89)90020-8

Nsiah, E. O. (2023). A Comparison of Variant Methods in Random Forest with Multicollinearity Data for Classification Prediction Modeling. *ProQuest Dissertations & Theses*. Available at: https://search.proquest.com

Substance Abuse and Mental Health Services Administration (SAMHSA). (2024). Mental Health Client-Level Data (MH-CLD). Available at: https://www.samhsa.gov/data/data-we-collect/mh-cld/datafiles

Su, C., Xu, Z., Pathak, J., et al. (2020). Deep learning in mental health outcome research: a scoping review. *Translational Psychiatry*, 10, 116. https://doi.org/10.1038/s41398-020-0780-3

Sung, M., Rees, V. W., Lee, H., & Jalali, M. S. (2024). Assessment of Epidemiological Data and Surveillance in Korea Substance Use Research: Insights and Future Directions. *Journal of Preventive Medicine*. Available at: https://ncbi.nlm.nih.gov

Verma, A., & Supekar, K. (2024). Novel Neural Network Models for Predicting Mental Health Outcomes in the US Youth Population. *Journal of Student Research*. Available at: https://jsr.org

Ware, O. D., Lee, K. A., Lombardi, B., & Buccino, D. L. (2024). Artificial Neural Network Analysis Examining Substance Use Problems Co-Occurring with Anxiety and Depressive Disorders Among Adults Receiving Mental Health Treatment. *Journal of Dual Diagnosis*, 1–12. https://doi.org/10.1080/15504263.2024.2357623

Ware, O. D., Lee, K. A., Lombardi, B., & Buccino, D. L. (2024). Artificial Neural Network Analysis Examining Substance Use Problems Co-Occurring with Anxiety and Depressive Disorders Among Adults Receiving Mental Health Treatment. *Journal of Dual Diagnosis, Taylor & Francis.*

Ware, O. D., Strickland, J. C., Smith, K. E., Blakey, S. M., & Dunn, K. E. (2023). Factors Associated with High-Risk Substance Use in Persons Receiving Psychiatric Treatment for a Primary Trauma- and Stressor-Related Disorder Diagnosis. *Journal of Dual Diagnosis*, 19(4), 199–208. https://doi.org/10.1080/15504263.2023.2260340

Ware, O. D., Zerden, L. D., Duron, J. F., & Xu, Y. (2024). Prevalence of Co-Occurring Conditions Among Youths Receiving Treatment with Primary Anxiety, ADHD, or Depressive Disorder Diagnoses. *Frontiers in Child and Adolescent Psychiatry*. Available at: https://frontiersin.org