Machine Learning and Psychological Disorders: An Analysis of Research on Machine Learning Diagnosis

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Claire Toussaint
Fall, 2020.

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Madhav V Marathe, Biocomplexity Institute

Machine Learning and Psychological Disorders

An Analysis of Research on Machine Learning Diagnosis

Claire Toussaint
Computer Science
University of Virginia
Charlottesville VA United States
ct4wa@virginia.ed

ABSTRACT

The current method for psychiatric diagnosis of mental illness is not as reliable as it should be. Instead of psychiatric evaluations by medical professionals, which can have many unreliable factors, machine learning could aid in diagnosing psychiatric disorders through neuro-imaging data and other evaluations. Previous research has been able to distinguish between patients with mental health conditions and patients without. However, until recently, research was not able to distinguish between different psychiatric disorders with similar symptoms.

The first analysis is of a study conducted to determine the importance of symptoms, cognition, and other multilevel variables for psychiatric disease classifications by machine learning. The Random Forest algorithm was used to generate the importance of the variables in diagnosis. The second analysis is of new research by the University of Tokyo. The research used machine learning to train six different algorithms to differentiate between MRI scans of patients who were either neurotypical, schizophrenic, or had autism spectrum disorder, and matched a psychiatrist's diagnosis with 85 percent accuracy. Lastly, a study on symptom differences of Major Depressive Disorder in psychiatric and general hospitals in China will be analyzed. Random Forest was used to create a predictive model of 62 variables in order to improve diagnostic accuracy.

With each study, analysis will identify questions, discuss limitations, and propose further research. The analysis will also discuss the importance of a secure database to store research data. The use of machine learning in correctly diagnosing psychiatric disorders could change the way we view mental disorders and subsequently progress treatment research and treatment options. Early and accurate detection of these illnesses could drastically improve or save lives.

INTRODUCTION

Mental illness affects nearly one in five Americans in a given year, with 50 percent of Americans being diagnosed with a mental illness or disorder in their lifetime. Both mental and physical health are key in overall health, and they are not independent of each other. Mental illness increases the risk of physical health problems, especially strokes, type 2 diabetes, and heart disease. Chronic health conditions also increase the risk of developing mental illness [1]. With the substantial number of people affected by mental illness, it is important that diagnosis is as accurate as possible to allow for the most effective treatment.

Currently, mental illness diagnosis is performed by a doctor or mental health professional. A physical exam or lab tests may also be used to rule out physical problems. In a psychological evaluation, the doctor or psychologist talks to the patient about symptoms, thoughts, feelings, and behavior patterns. A questionnaire may also be used to aid diagnosis [2].

Although the current way to diagnose mental illness is the best available, it is not highly accurate. There are several factors that influence the reliability of current diagnosis models, such as the patient's psychological state, clinician inconsistency, and inadequacies in defining different illnesses and recognizing different symptoms for the same illness [3]. A methodical, unbiased approach for diagnosis would help reduce these inaccuracies and improve accuracy.

BACKGROUND

1 Machine Learning

Machine learning creates algorithms that learn from data to improve their accuracy over time. These algorithms are trained with large amounts of data to find features and patterns so that they can make predictions on new data. The algorithms discussed in these studies are classifiers, which is categorized under supervised learning. Supervised learning has an input variable and an output variable, and tries to create an algorithm to accurately map new data. Classifiers use the input labels to predict the label, or class of new, unlabeled data points [4].

2 Psychiatric Disorders

Psychological conditions: Psychiatric disorders, or mental health disorders, entail a wide range of mental health conditions. These conditions affect mood, thinking, and behavior. Disorders are thought to be caused by genetic and environmental factors, such as inherited traits, environmental exposures before birth, and brain chemistry. Certain genes may increase the likelihood of developing a mental disorder, thus it is more common in people who have blood relatives with a disorder. Development of disorders can also be triggered by life events or environmental exposures before birth, like alcohol or drugs. Also, when there is trauma to the brain and neural networks are impaired, it can lead to depression and other emotional disorders [2].

RELATED WORK

There are two main application areas of machine learning and mental health where current research falls into. The first is detecting, understanding, and diagnosing mental health symptoms. The second is assessing patient-clinician relationships and improving mental health treatment.

Applications that aim to detect mental health symptoms with machine learning are achieved by a wide variety of methods. There are studies that use acoustic features of speech, Twitter tweets, and mobile sensing data like typing dynamics and analysis of text messages to detect symptoms. Other studies aim to predict future suicide risks by analyzing suicide notes and suicidal periods evident from text messages. Studies were also conducted to predict episodes of mania or depression in individuals with bipolar disorder. Social media analysis has been performed to analyze characteristics of content shared online relating to mental health and which types of comments should warrant a response by moderators.

In assessing patient-clinician relationships, machine learning studies have analysed questionnaire data about patient experiences with their doctor in a psychiatric hospital. Studies have also been conducted to try to improve treatment delivery. For example, a study applied machine learning in a mobile application to recommend personalized coping strategies by learning from the users' engagement with different stress interventions [5].

SYSTEM DESIGN

The analysis of machine learning with psychological diagnosis will follow three central questions. These questions are: what factors are important for use in machine learning diagnosis, how can we utilize machine learning for diagnosis, and how can we improve diagnosis accuracy? The study of "Relative importance of symptoms, cognition, and other multilevel variables for psychiatric disease classifications by machine learning" will be analyzed to answer the first question. This paper highlights the importance of taking a holistic approach to analyze disorders. It was chosen to emphasize that many variables need to be utilized to create an unbiased and accurate application. To answer the second question, the study "Machine-learning classification using neuroimaging data in schizophrenia, autism, ultra-high risk and first-episode psychosis" will be analyzed. While there are many different approaches to using machine learning as a diagnostic tool, this study was chosen because it has unbiased, measurable data from its use of neuroimaging. Studies discussed in the previous section potentially have more user bias in their analysis. Lastly, the third question will be analyzed with the study "Symptomatology differences of major depression in psychiatric versus general hospitals: A machine learning approach." This study was chosen for its inclusion of sociological factors in psychological diagnosis. It also highlights how diagnosis may differ under different environmental and circumstances. These factors need to be considered and translated into a machine learning application for an accurate diagnosis.

1 Importance of Symptoms

The first question will be analyzed with the study "Relative importance of symptoms, cognition, and other multilevel variables for psychiatric disease classifications by machine learning." This study was conducted to determine the importance of symptoms, cognition, and other multilevel variables for psychiatric disease classifications by machine learning. The study used machine learning to estimate the relative importance of different data for classifying cases into categories of schizophrenia, schizoaffective disorder, bipolar disorder, unipolar depression, and healthy controls. A second aim of the study was to analyze individual features of a healthy control misclassified as having schizophrenia, again with machine learning, to determine which data may be the most relevant for classifying schizophrenia.

The study utilized data from 113 psychiatric cases and 51 healthy participants from two public hospitals in New York City. Each participant was interviewed by a clinician with a master's degree or higher to assess psychiatric symptoms and disorders, and other information was gathered. This included parental ages at birth, Global Assessment of Function (GAF), Positive and Negative Syndrome Scale (PANSS), Verbal Comprehension (VCI), Working Memory (WMI), Processing Speed (PSI), and Perceptual Organization (POI).

Four machine learning techniques were assessed for preliminary efficacy in classifying psychiatric cases versus healthy cases. Random Forest (RF), Linear Discriminant Analysis (LDA), Adaptive Boosting (AdaBoost), and Support Vector Machine (SVM) were assessed, and RF was found to be the best technique (See Appendix). Three different groupings were created, and the mean decrease in accuracy (MDA) was analyzed. For analyzing the healthy controls misclassified as psychiatric cases, the healthy controls were split into three groups for testing and training combinations. Each control that was misclassified was individually assessed for variables consistent with schizophrenia cases [6].

1.1 Results

While the extensive results are not essential in the analysis, is it important to note that different psychiatric cases were able to be separated based on the data. The most important data in separating schizophrenia/schizoaffective from bipolar/unipolar depression was GAF, while schizophrenia was separated from all other psychiatric disorders by low GAF and paternal age. Misclassified healthy cases showed lower nonverbal abilities and mild negative psychopathology symptoms [6].

These results show the importance of utilizing a wide range of data in classifying different psychiatric disorders. This approach can be used to further improve both the accuracy of machine learning diagnosis and determining what factors are important for use in machine learning diagnosis of a particular disorder.

1.2 Limitations

The study did not have an even distribution of diagnoses and sex of participants, which could have affected the results. Also, additional data such as life exposures and genetic history of disorders could be insightful for future studies. The number of participants was fairly low, thus overfitting may have occurred with the machine learning analysis. Future trials or studies should utilize more participants to avoid this.

1.3 Questions

This study essentially validates the current model of psychiatric diagnosis as it makes a decision based on assessment of symptom and cognition data. At what point does analysis of symptoms using machine learning become more effective and accurate than psychiatric evaluation? Would a machine learning analysis to find the importance of symptoms utilized by a medical professional be more beneficial than a medical professional determining the most important symptoms to be used by machine learning in diagnosing a disorder? If the goal is to have medical professionals eliminated from the picture entirely, solid trust must first be established in machine learning diagnosis for societal adoption.

1.4 Future Research

Future studies should be conducted to discern between more types of mental disorders. While other disorders may be easier to differentiate between and may not need comparison, it is possible that unknown links between them could be discovered. Additional data could also provide further insights into the patterns and most prevalent symptoms. Also, to limit any overfitting with the machine learning analysis, more data should be added. This data could either be additional types of mental disorders or additional participants with the disorders already analyzed in this study.

2 Machine Learning Classification

The possible utilization of machine learning to diagnose psychiatric disorders will be analyzed with the study “Machine-learning classification using neuroimaging data in schizophrenia, autism, ultra-high risk and first-episode psychosis.” This study was conducted with four aims. The first aim was to construct and compare machine learning classifiers to predict cases such as schizophrenia, autism spectrum disorder (ASD), and typically developing (TD) based on their MRI scans. The second aim was to determine the most important brain feature groups for classification, namely cortical thickness, surface area, and subcortical volume. To analyze the classifiers for accuracy, the third aim was to assess the consistency of the classifiers with clinical severity. Lastly, the fourth aim was to predict the diagnosis category of ultra-high risk for psychosis (UHR) and first-episode psychosis (FEP) subjects using the trained classifiers.

To complete the study, the data of 131 schizophrenic spectrum, 45 high functioning ASD, and 125 TD participants were recorded. After assessing MRI-scan images for quality and accuracy, only 97 schizophrenia spectrum, 36 ASD, and 106 TD scans were analyzed. Of the 97 schizophrenia spectrum participants, 26 had UHR, 17 had FEP. The age ranges for each disorder varied, and individuals with ASD were all males whereas the rest of the individuals were of mixed sex. The data was thoroughly processed, with any unfixable abnormalities discarded from the study.

In order to avoid bias by selecting one classifier, six classifiers were analyzed. The classifiers were Logistic Regression (LR), Support Vector Machine, Random Forest, AdaBoost, Decision Tree (DT), and k-Nearest Neighbor (kNN) (See Appendix). With each classifier, there was one multiclass classification (schizophrenia/ASD/TD) and three binary classifications (schizophrenia/ASD, ASD/TD, and

schizophrenia/TD). The accuracy in classification was not the only accuracy metric to analyze each classifier. The confusion matrix, recall score, precision score, and F1/F2 scores were also calculated. The top performing multiclass and binary class classifiers were then used to test 26 individuals with UHR and 17 with FEP [3].

2.1 Results

While specific results are not important for analysis, the findings for the first two aims of the study were that SVM and LR were the best performing classifiers, and cortical thickness and subcortical volume were the most useful feature groups. Regarding the third aim of the study, LR, SVM, and DT's output were consistent with the individual's clinical severity. Lastly, the category of both UHR and FEP individuals were correctly predicted by the trained classifiers [3]. These results show us that machine learning is capable of accurately predicting disorders based on objective data.

2.2 Limitations

While the study performed meticulous data collection, processing, and analysis, there were several flaws. All of the schizophrenia individuals and several of the ASD, UHR, and FEP individuals were medicated, which could have effects on the results. Also, as previously mentioned, all of the ASD individuals were male. While the ASD classifications did not seem to be affected by this, future studies should consider using data from both males and females for all groups.

2.3 Questions

While this study provides substantial findings in using MRI scans and machine learning to diagnose schizophrenia and ASD with certain feature groups, the accuracies do not surpass 85% for any combination of classifier or feature. An ethical question is raised with how much weight the machine learning prediction should be given in actually diagnosing a patient compared to the weight of the diagnosis by a psychiatrist. Unlike the first study which validates symptoms recorded by a medical professional, this study uses scientific data from the MRI to make predictions. Would this type of diagnosis be better accepted by society? At what accuracy will machine learning diagnosis by MRI be trusted more than a medical professional's diagnosis by symptom analysis?

2.4 Future Research

Since this study provides promising evidence on the ability to diagnose schizophrenia and ASD using machine learning, future research should assess other possible disorders that can be accurately quantified and analyzed from brain imaging through machine learning, such as bipolar disorder.

3 Diagnosis Accuracy

To analyze the third question pertaining to increasing the accuracy of machine learning diagnosis, the study "Symptomatology differences of major depression in psychiatric versus general hospitals: A machine learning approach" will be analyzed. Major depressive disorder (MDD) is highly prevalent in China, yet the diagnosis and treatment rates are disproportionately low. Mental-health services are mostly hospital-based, and patients visit either general hospitals or psychiatric hospitals for diagnosis and treatment. However, a survey in a general hospital in Shanghai found that only 18.5% of patients with MDD were diagnosed with it. This is a result of a combination of factors, such as general-hospital patients displaying atypical symptoms and the lack of skilled mental health professionals.

The aims of the study were to identify predictors of patient choice of mental health services and identify the differences in symptomatology of patients with MDD in general hospitals and psychiatric hospitals. Using machine learning to obtain these results, the goal was to improve the accuracy of clinical diagnosis and provide references for establishing health policy.

The study used data from the National Survey on Symptomatology of Depression (NSSD) from 16 general hospitals and 16 psychiatric hospitals in 22 cities in mainland China. The 1500 patients in the study consisted of first-visit patients of age 18 or older who met the criteria for MDD diagnosis and had no other psychiatric disorders. A 62 symptom questionnaire was created and given to the participants to complete while in the same room as a clinician. The symptoms on the questionnaire were in random order to minimize systemic bias, and most of the questions were on a four-point Likert scale. The questions were broken down into the following categories: eight Socio-demographic variables, five causal attribution variables, forty-nine symptoms variables, eleven symptoms of specifiers for depressive disorders in DSM-5, the Diagnostic and Statistical Manual of Mental Disorders, and 24 non-DSM symptoms. For machine learning analysis, these variables were classified into either social-demographic part, causal attribution part, and symptoms.

The importance of variables in predicting MDD was calculated by using the Random Forest machine learning technique. With each variable, there are two measures of importance in RF. The first measures the mean decrease in accuracy when the variable is excluded. The second measure is the Gini impurity, which measures the likelihood of an incorrect classification when a variable is chosen to split the data. However, only the decrease in accuracy was considered for this study. Ten-fold cross

validation was used on the training and testing sets, and the overall accuracy of the random forest was 73.4% with an out-of-bag (OOB) error of 26.60%. The model included 500 decision trees [7].

3.1 Results

Following the first aim of the study, RF found that particular symptoms of depression are strong predictors of patient choice of mental health facilities. Regarding the second aim, RF found that general hospital patients had higher frequencies of suicidal ideation, psychosis, weight change, hypersomnia, and a tendency of denying emotional/cognitive symptoms compared with psychiatric hospitals patients [7].

3.2 Limitations

Unlike the previous studies, only one algorithm was used. While RF is proven to be effective and accurate for classifying large amounts of data, as it forms a strong classifier from many weak classifiers, choosing only one technique is a flaw. It creates bias that should be avoided by utilizing multiple techniques and selecting the best one, even if just preliminary training and testing is done. The study may also have flaws in the data due to the stigma surrounding mental health in China. While it was mentioned that the clinician would be present in the room to answer questions, it was not mentioned if the results were anonymized. The participants may be less willing to answer truthfully knowing that their results will be read by the clinician if they must also provide their name. These factors may give insight as to why the accuracy is fairly low.

3.3 Questions

This study is particular to hospitals in mainland China, where the health system and stigma around mental health differs from that of the United States. It is not guaranteed that this study could map well to a different population. Studies that seek to reproduce this study will need to take into consideration social factors present for the tested population, and may need to change the questionnaire and aims accordingly. While this study ultimately aims to increase clinical diagnosis accuracy, the ethical question of acceptance is again brought up. Machine learning techniques to find the most important symptoms in diagnosis, which can then be used by medical professionals when making a diagnosis, may be more readily accepted. Compared to the previous study, this study depends more on behavioral analysis. While it may not significantly reduce the bias in the current psychological diagnosis model, it is a step in the right direction.

3.4 Future Research

While this study finds that symptoms are important in predicting health-seeking behavior, it does not conclude why. A possible reason could be that there is a higher level of stigma in general hospitals that may alter the way a patient is expressing his or her symptoms. Another possible explanation could be that certain symptoms predict the choice of mental health care, or facility, and thus different proportions of subtypes of depression are present in hospitals. While these questions may not be answered with machine learning, answers could better support and explain the findings in this study.

4 Database Systems

Developing technological systems to analyze and diagnose mental health disorders brings socio-technical challenges. Mental disorders already carry a stigma, and thus putting them through a scientific lens amplifies this, especially when there are concerns about the reliability and accuracy of modeling within the scientific community. In order to create reliable algorithms, the data must be as inclusive as possible since mental disorders affect a broad range of demographics. If a model is using data of a particular demographic, it may be highly accurate within the model but not scale well to other demographics. This creates a biased model that may not give helpful insights or even give an incorrect analysis. With every recommendation for data collection, concerns with privacy may arise since the data is sensitive. Secure and reliable database systems need to be used to collect and store the data that is used by the machine learning models.

Database security is crucial for databases storing identifiable personal data. Concerns with machine learning and mental health diagnosis may include the recording of personal data, thus security needs to be enforced. This will ease individuals into consenting to their data being recorded for use in a study. Security is also important in ensuring data is not mistakenly edited or deleted. Access control should be implemented on each database to restrict system access to authorized users. Ideally, Role-Based Access Control (RBAC) would be used. This would allow individuals to be given a role, in which the role determines their privileges. Roles should be limited to the least amount of privileges necessary to perform their task. Without accurate data, the machine learning algorithms do not amount to anything.

CONCLUSION

The acceptance of mental disorders as a medical condition has opened the doors to a more scientific approach in analysis. Machine learning has been used to do symptom analysis and predictions of mental disorders with a wide range of techniques from MRI scans and questionnaires to mobile phone sensing. The current model of psychiatric

diagnosis has too many factors that affect its reliability, and the addition of machine learning to predict which symptoms are the most important in distinguishing between disorders could greatly aid accuracy. The replacement of the current model for predicting certain disorders by MRI scans may be more accurate but face potential social opposition. Studies pertaining to machine learning and psychiatric disorders could provide crucial changes to how society perceives mental disorders, the accuracy in diagnosis, and the effectiveness of treatment.

REFERENCES

[1] Anon, 2020. Learn About Mental Health. Cdc.gov.

[2] Mayo Clinic Staff, 2020. Mental illness - Diagnosis and treatment. Mayoclinic.org.

[3] Yassin, W. et al., 2020. Machine-learning classification using neuroimaging data in schizophrenia, autism, ultra-high risk and first-episode psychosis. Translational Psychiatry, 10(1).

[4] IBM Cloud Education, 2020. What is Machine Learning?. Ibm.com.

[5] Thieme, A., Belgrave, D. and Doherty, G., 2020. Machine Learning in Mental Health. ACM Transactions on Computer-Human Interaction, 27(5), pp.1-53.

[6] Walsh-Messinger, J., Jiang, H., Lee, H., Rothman, K., Ahn, H. and Malaspina, D., 2019. Relative importance of symptoms, cognition, and other multilevel variables for psychiatric disease classifications by machine learning. Psychiatry Research, 278, pp.27-34.

[7] Cui, L. et al., 2020. Symptomatology differences of major depression in psychiatric versus general hospitals: A machine learning approach. Journal of Affective Disorders, 260, pp.349-360.

APPENDIX

Adaptive Boosting (AdaBoost)

AdaBoost is an ensemble method that sequentially trains predictors in the attempt to correct its predecessor. AdaBoost first trains a base classifier, such as a Decision Tree. It selects the misclassified predictions from the base and gives them more weight. It then trains another classifier with the new weights, and repeats the process. Combining weak classifiers creates a strong classifier, as each iteration focuses on correctly classifying the previous misclassified instances.

Decision Tree (DT)

Decision Trees are used for both classification and/or regression. DTs are also the fundamental component of Random Forests. A decision tree is a structure that starts at a root node and is partitioned into subsequent nodes until there is a leaf node, which has no more partitions. DTs are trained using Information Gain (IG) to determine which feature is most useful in discriminating between classes, thus the order of the tree. At each node, the feature with the highest IG is used as the splitting feature. The final features, or leaf nodes, are the classifications.

K-nearest neighbors (kNN)

K-nearest neighbor is used for both classification and/or regression. It assumes similarities between new instances and existing instances by assuming similarities follow from closeness. When there is a new instance to be classified, the euclidean distances between the instance and its k nearest neighbors are found. Of the k-nearest neighbors, the number of occurrences of each classification is counted, and the instance is assigned to the classification with the highest count.

Linear Discriminant analysis (LDA)

LDA is a classification and dimensionality reduction technique, which projects features in a high dimensionality space onto a lower dimensionality space. It can be useful when multiple classes are involved, as Logistic Regression is a binary classification. LDA calculates the probability of an instance belonging to each class, then classifies the instance to the class with the highest probability.

Logistic Regression (LR)

Logistic Regression is used to calculate the probability that an instance belongs to a particular class. It is a binary classifier that predicts a 1, if the probability is greater than 50%, meaning it belongs to that class, or a 0 if it does not belong. The probabilities are found by first calculating the weighted sum of the input features, plus a bias term. The logistic of the result is then found, which is the probability.

Random Forest (RF)

Random Forests are used for classification and/or regression. It is part of ensemble learning, which aggregates the prediction of a group of predictors. RFs obtain predictions from multiple Decision Trees and use the prediction with the most votes. Typically, trees are trained with the same algorithm but utilize different subsets of the data.

Support Vector Machine (SVM)

SVM is used for binary classification and/or regression. In classification, SVM finds a hyper-plane that creates a boundary between the data. The goal is to maximize the distance between the two data sets to the hyperplane, or the decision boundary. The data points closest to the decision boundary on each side are used to create support vectors, which run parallel to the decision boundary. The distance between the support vectors and decision boundary should be maximized while limiting margin violations. The margin violations are instances between the support vectors and the boundary or on the incorrect side of classification.