Predicting Mental Health using Machine Learning Algorithm

Tanu Maurya, Abhinav Akash Tripathi, Himanshu Kushwaha

Department of Computer Science & Engineering in Data Science,

ABES Engineering College, Ghaziabad, U.P.

E-mail: Tanu.20B1541058@abes.ac.in ,

Abhinav.20B1541027@abes.ac.in ,

Himanshu.20B1541014@abes.ac.in

ABSTRACT: The increase of mental health problems and the need for effective medical health care have led to an investigation of machine learning that can be applied in mental health problems. Mental health prediction is one of the most essential parts of reducing the probability of serious mental illness. Meanwhile, mental health prediction can provide a theoretical basis for public health department to work out psychological intervention plans for medical workers. Our aim is to create a machine learning model that can predict mental health or the possibilities of a person suffering from mental health issues based on the data set. The final outcome will be a mental health prediction website that will help to predict weather the person is suffering from mental health issues based on the data set.

INTRODUCTION:

A person's mental health is determined by both their current state of mind and how they are interacting with the world around them. Mental illness is brought on by anomalies in the brain's chemistry. A person's level of mental health acts as a gauge for how to treat their illnesses effectively. It is crucial to monitor the mental health characteristics of various groups in order to anticipate any health-related anomalies. There are working adults, college students, and high school students living in the neighbourhood. It's a common misconception that stress and unhappiness affect people of all ages and socioeconomic backgrounds. It is essential to assess the mental health of various categories at various points in time in order to prevent major sickness. Healthcare professionals will soon be expected to take a patient's

mental health profile into account in order to administer better treatment and hasten their recovery.
Some of the most serious mental health conditions, like chronic illnesses, bipolar disorder, and schizophrenia, develop gradually over time and have early-stage signs that can be identified. Such disorders might be prevented or better managed. If anomalous mental states are identified early on in the disease's progression, additional attention and therapy can be given. Therefore, making assumptions about someone's mental state based on how they act or look is a sophisticated psychological science that has not yet been mechanised. Although there are screening test solutions, they are not practical for large populations due to time and financial constraints. Furthermore, diagnosis-based methods unintentionally discourage people who are ill from participating.
Depression and anxiety are significant conditions that have a global impact on people's health. Men and women of all ages, including youngsters and the elderly, are affected by them. The impact of anxiety and depressive disorders on health and well-being are extensive. Different somatic symptoms like gastritis, acid reflux, palpitations, insomnia or hypersomnia, tremors, significant weight loss or gain, and different psychosocial manifestations like low mood, social withdrawal, decreased workplace productivity, suicidal ideation or attempt, and lack of concentration are all caused by them. A number of additional lifestyle problems, including ischemic heart disease, hypertension, diabetes, unintentional accidents, and purposeful harm, are significantly increased by depression and anxiety. Depression has a close relationship with suicidal thoughts, and depression itself can result in suicide. They suffer from a variety of contagious illnesses, including HIV and tuberculosis. People who experience depression and anxiety are usually stigmatised by society and socially excluded by their family. They could do poorly at job and in educational institutions. People are thereby being denied access to social and economic opportunities, which lowers their quality of life. Economic stress is a pervasive and frequently intangible manifestation that fuels a cycle of poverty and ill health. Low- and middle-income households make up the bulk of those impacted.

Smartphones, social media, neuroimaging, and wearable technology have made it possible for medical professionals and mental health researchers to obtain a tonne of data quickly. The ability to analyse these data with machine learning has grown. Advanced probabilistic and statistical methods are used in machine learning to build computers that can independently learn from data. As a result, data patterns can be

Methods	Accuracy (%)
Logistic Regression	79.63
KNeighbors Classifier	80.42
Decision Tree classifier	80.69
Random Forests	81.22
Stacking	81.75

Figure 1. Accuracy of all Classifiers

easier to find and correctly identify, as well as more precise forecasts from data sources. Machine learning has benefited a variety of fields, including natural language processing, speech recognition, computer vision, and artificial intelligence. This is because it enables programmers and researchers to extract important information from datasets, provide individualised experiences, and create intelligent systems. By enabling speedy and scalable analysis of complex data in fields like biology, ML has significantly helped advancement. Similar analytical techniques are also being used to study mental health data, with the aim of enhancing patient outcomes and enhancing knowledge of psychological disorders and their treatment.

MATERIALS AND METHODS:

Data Preparation:

This study makes predictions on the mental health of Chinese healthcare professionals during COVID-19 using data from a survey titled "Mental Health Status of Medical Workers During COVID-19," which was conducted in Changchun, Jilin Province, China, from June 1 to June 7, 2020. The survey's participants were medical professionals who took part in epidemic control and prevention. We chose 150 of the city of Changchun's 220 grassroots medical units based on the population's health and the features of the region's distribution, and we then randomly picked 35 medical professionals for each of the units.

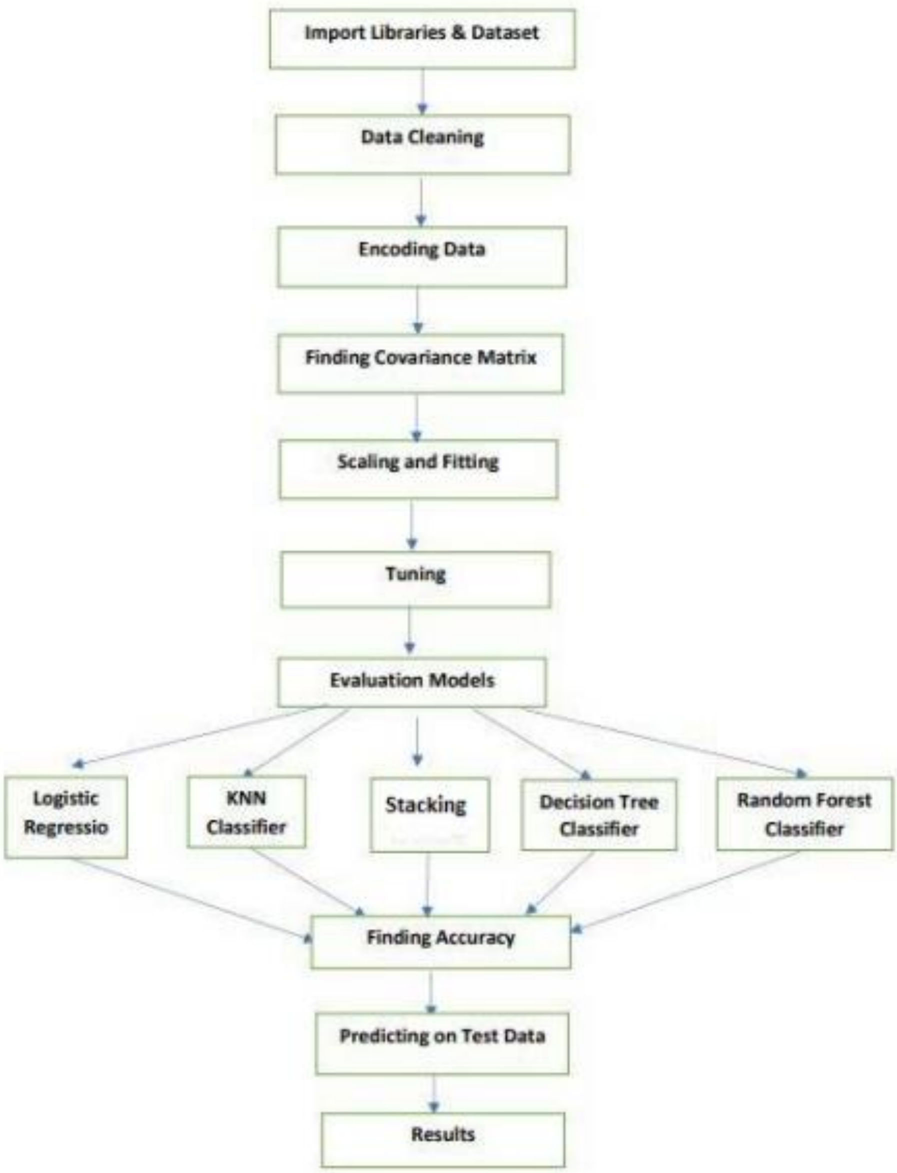


Figure 2. Flow Chart

Data collection, data cleaning, encoding, finding covariance matrices, scaling and fitting, tuning, evaluating models, finding accuracy, and predicting data and results are all included in the knowledge discovery from data process as shown in Figure 2. We start by taking into account the dataset with 1259 entries and 27 columns. The next step is data cleaning, which entails identifying any inaccurate, incomplete, unnecessary, or missing data and then updating, replacing, or removing it as necessary. Three columns contain the missing data, as we discovered. The special value Not a Number, or NaN, is used to denote a cell with no value in Data Frames and Numpy arrays. Data encoding is the following step. We use this categorical data encoding approach when the categorical feature is designated as ordinal.

It is crucial to keep the order in this situation. In order to reflect the sequence, the encoding should do so. During label encoding, each label will be converted into a value of an integer. We'll then search for the covariance matrix after that. It is one of the key matrices in data science and machine learning. It provides details about feature co-movement (correlation). The variances of the variables will be in the main diagonal of the matrix of variance-covariance, and the covariances will be between each pair of variables in the other matrix locations. The means of each variable are contained in the mean vector.

In feature scaling, we place the independent features of the data into a predetermined range. During data pre-processing, it manages drastically changing values, units, or magnitudes. The dataset was then divided into a training and a testing data set. The importance of a feature comes next. The fundamental technique of directing variable usage to what is most practical and efficient for a certain machine learning system makes feature selection crucial in machine learning. Tuning is the following phase. Enhancing a model's performance while preventing overfitting or excessive variation is the process of tuning. In machine learning, this is accomplished by choosing the proper hyperparameters. Hyperparameters can be compared to the "dial" or "knobs" of a machine learning model. The models are then evaluated using a variety of machine learning methods, such as stacking, logistic regression, K-nearest neighbor classifier, decision tree classifier, and random forest classifier.

Logistic Regression

A well-known machine learning algorithm that falls within the supervised learning methodology is logistic regression. By using a set of unbiased variables, we can use this strategy to predict a certain dependent variable. To predict a particular structured variable's output, one uses logistic regression. So

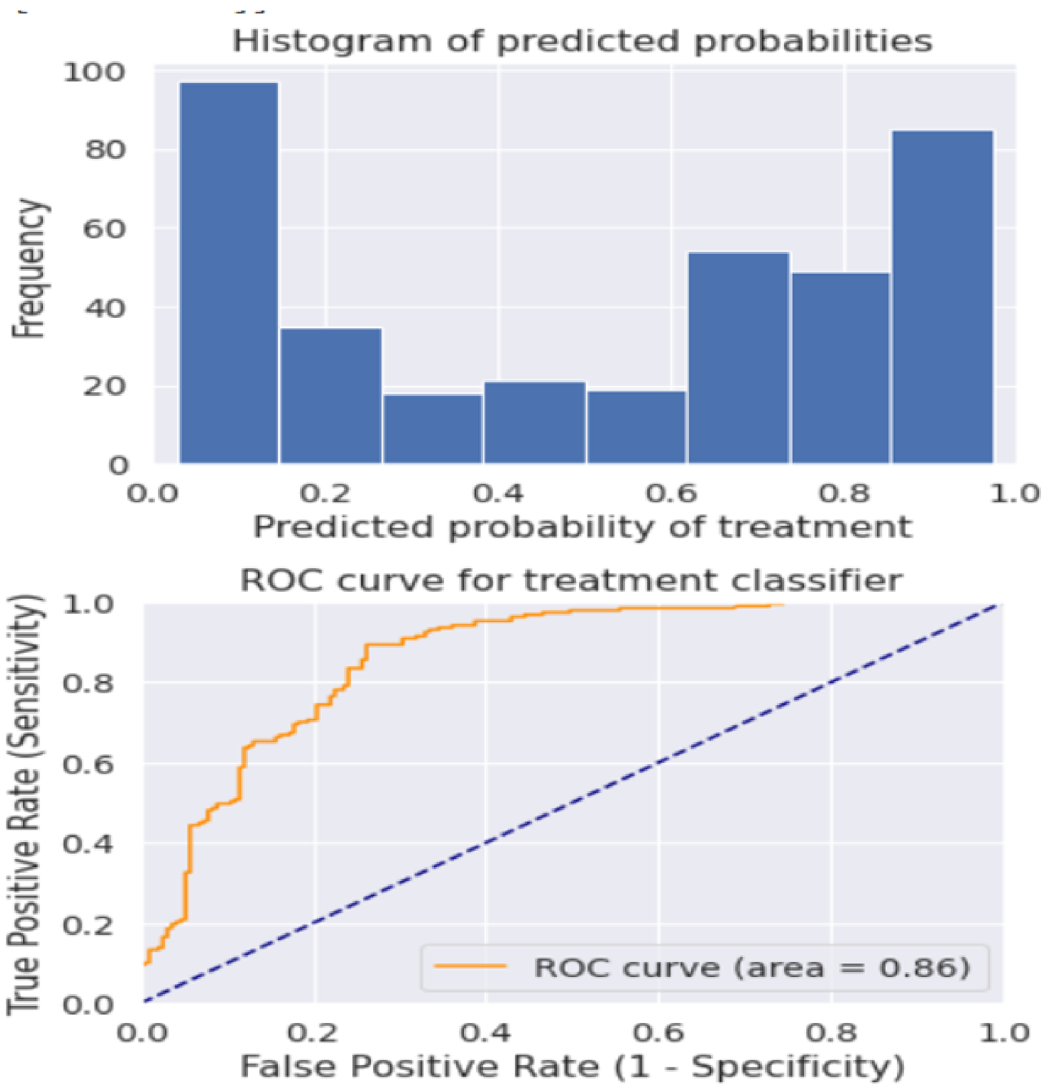


Figure 3.

the result should be a categorical or discrete value. It can be 0 or 1, Yes or No, true or false, and so on, but it delivers probabilistic values that are somewhere between 0 and 1 instead of giving exact values like 0 and 1.

K nearest neighbour classifier

A fundamental machine learning algorithm that uses the Supervised Learning method is called the K-Nearest Neighbour. Existing cases and fresh case/data will be comparable in the K-NN approach. KNN is a non-parametric algorithm that does not assume anything about the distribution or the emphasised data. Additionally, it functions with several classes.

Decision tree classifier

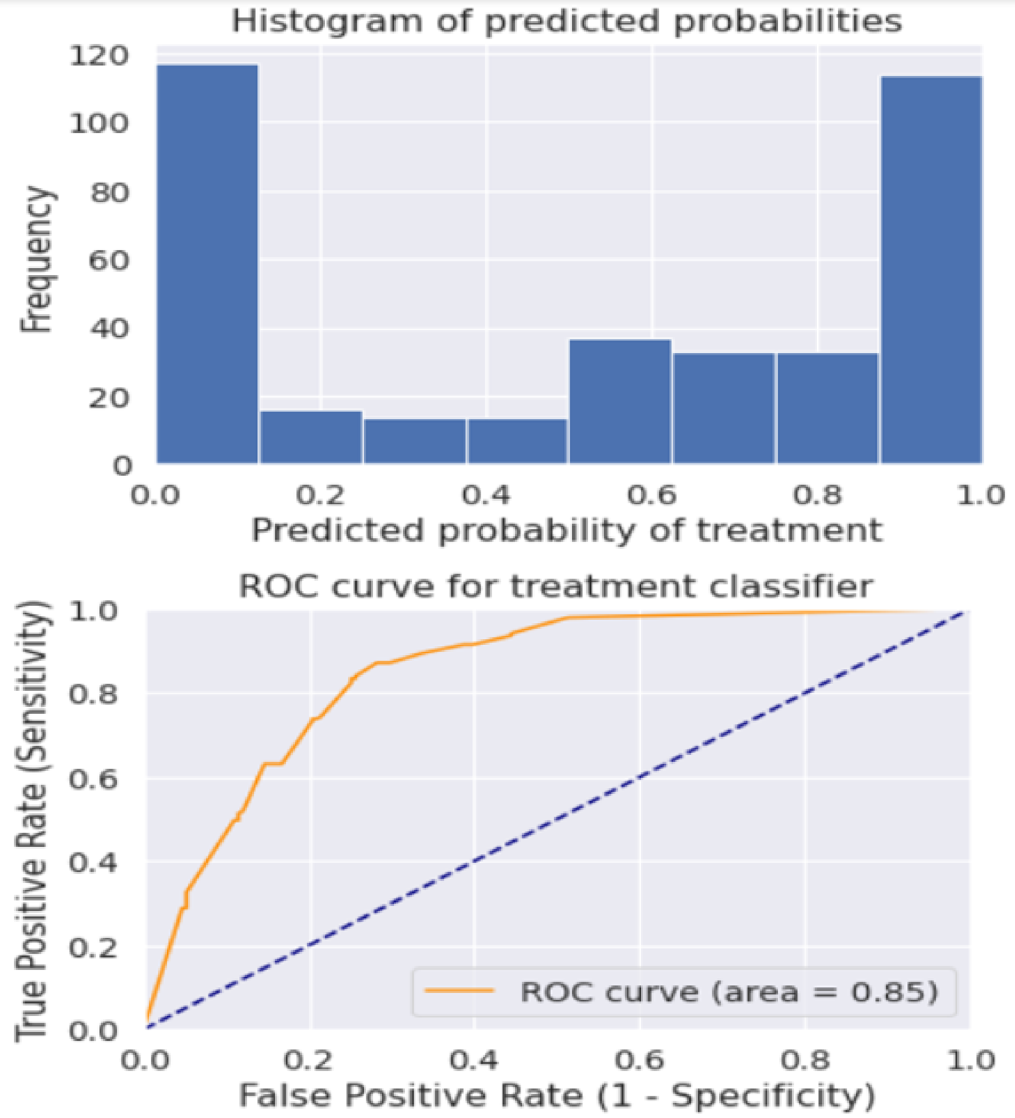
The most popular supervised machine learning method used in data mining is the decision tree. Using a decision tree is a visual representation of a statistical likelihood or the order of events, actions, or consequences.

Random forest classifier

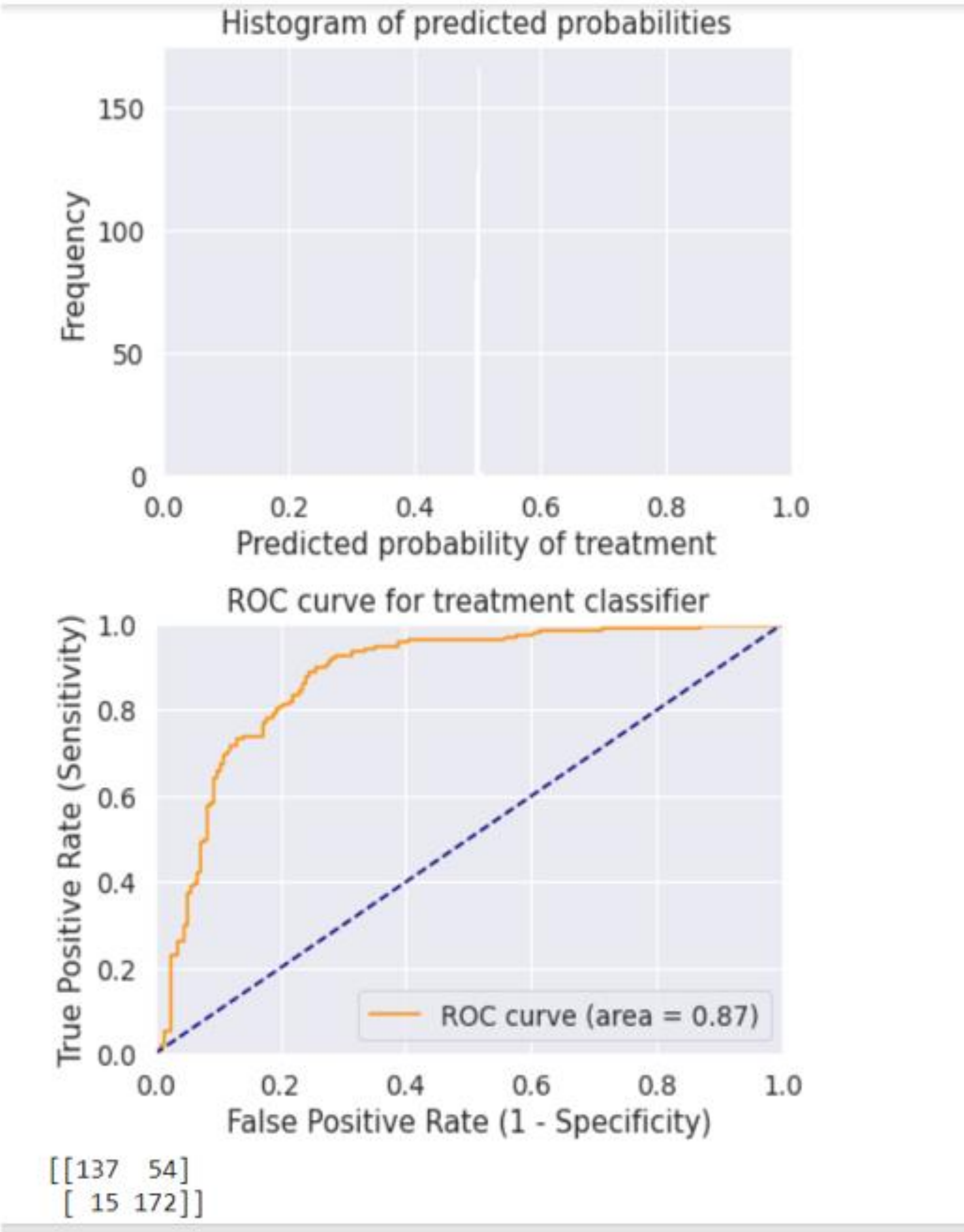
Using a method called random forest, which is based on supervised machine learning, classification and regression issues can be solved. However, classification is a common usage for it. It is known as a random forest because it aggregates numerous decision trees to form a "forest" and feeds random features from the input dataset to them.

Stacking

A machine learning ensemble algorithm called "Stacking" stands for stacked generalisation. It involves combining predictions from various machine learning models on the same dataset, similar to bagging and boosting.




```
[[142  49]
 [ 30 157]]
```



Results

This study found five machine learning methods: random forest, decision tree and stacking, logistic regression, and k closest neighbour classifier. And we evaluated how well they were able to spot mental health problems. We first ran the classifiers with all 27 attributes that were extracted from the text documents, and then we ran them with 8 additional attributes that were chosen using the feature selection approach. The Accuracy of a given test set for a classifier is the percentage of test set instances that are classified correctly by using the classifier. Figure 2 shows the accuracy of stacking is more accurate compared to other classifiers. The accuracy of any classifier will depend upon how well the classifier will classify the data set which is being tested. We measured that by using the area under the Receiver

Operating Curve. In the ROC area, a perfect test will represent an area of 1 and a worthless test will represent an area of 0.5. Figure 3-7 illustrates the graph of five classifiers on ROC Area values. We observed that the classifiers were more accurate in predicting the condition of mental health than other classifiers because the ROC area of all classifiers used is between 0.8 and 0.9.

Mental Health Prediction Website All Predictions

Name	Country	State	Tech Company	Student	Physical Health	Age	Sex	Family_history	Predictions
Vaishali	India	Madhya Pradesh	Yes	No	Yes	37	0	1	['Yes']
Renu Tripathi	India	Uttar Pradesh	No	No	Yes	50	0	1	['No']
Abhijat	India	Uttar Pradesh	No	Jimms	Sleep disk	20	1	1	['No']
John	India	Uttar Pradesh	Yes	No	Ok	34	1	1	['Yes']

The final prediction consists of 0 and 1. 0 means the person is not needed any mental health treatment and 1 means the person is needed mental health treatment.

Conclusion

Since there are numerous machine learning approaches accessible, it is crucial to compare them all and then choose the one that best fits the target domain. Today, there are numerous specialised programmes in the medical profession that can forecast disease quite accurately in advance, allowing for effective and quick therapy. In the proposed work, five different machine learning approaches that were employed to categorise a dataset of different mental health issues were compared. The results make it quite evident that all five machine learning methods produce more accurate outcomes. All classifiers have accuracy rates of more than 79%. The data set used in the research is very minimal and in the future, a large data set can be used and the research can be applied on the same for more accuracy.

REFERENCES

1. G. Miner, L. Miner, M. Goldstein et al., *Practical Predictive Analytics and Decisioning Systems for Medicine: Informatics Accuracy and Cost-Effectiveness for Healthcare Administration and Delivery Including Medical Research*, Academic Press, Inc., Cambridge, MA, USA, 1st edition, 2016.
2. M. Hamilton, "Development of a rating scale for primary depressive illness," *British Journal of Social and Clinical Psychology*, vol. 6, no. 4, pp. 278–296, 1967.
   View at: Publisher Site | Google Scholar
3. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, American Psychiatric Association, Reston, VA, USA, 2013.
4. M. I. Jordan and T. M. Mitchell, "Machine learning: trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.
   View at: Publisher Site | Google Scholar
5. F. Dabek and J. J. Caban, "A neural network based model for predicting psychological conditions," in *Brain Informatics and Health*, pp. 252–261, Springer International Publishing, Berlin, Germany, 2015.
   View at: Publisher Site | Google Scholar
6. G. Cho, J. Yim, Y. Choi, J. Ko, and S.-H. Lee, "Review of machine learning algorithms for diagnosing mental illness," *Psychiatry Investigation*, vol. 16, no. 4, pp. 262–269, 2019.
   View at: Publisher Site | Google Scholar
7. E. G. Pintelas, T. Kotsilieris, I. E. Livieris, and P. Pintelas, "A review of machine learning prediction methods for anxiety disorders," in *Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion—DSAI 2018*, Thessaloniki, Greece, 2018.
   View at: Publisher Site | Google Scholar
8. R. B. Rutledge, A. M. Chekroud, and Q. J. Huys, "Machine learning and big data in psychiatry: toward clinical applications," *Current Opinion in Neurobiology*, vol. 55, pp. 152–159, 2019.
   View at: Publisher Site | Google Scholar
9. 130 Executive Board, "Global burden of mental disorders and the need for a comprehensive, coordinated response from health and social workers at the country level," 2012, https://apps.who.int/iris/handle/10665/23741.
   View at: Google Scholar
10. M. Hadzic, M. Chen, and T. S. Dillon, "Towards the mental health ontology," in *Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine*, Philadelphia, PA, USA, 2008.
   View at: Publisher Site | Google Scholar

11. J. Diederich, A. Al-Ajmi, and P. Yellowlees, “Ex-ray: data mining and mental health,” *Applied Soft Computing*, vol. 7, no. 3, pp. 923–928, 2007.
View at: Publisher Site | Google Scholar
12. P. D. McGorry, A. R. Yung, L. J. Phillips et al., “Randomized controlled trial of interventions designed to reduce the risk of progression to first-episode psychosis in a clinical sample with subthreshold symptoms,” *Archives of General Psychiatry*, vol. 59, no. 10, pp. 921–928, 2002.
View at: Publisher Site | Google Scholar
13. T. I. Mueller, A. C. Leon, M. B. Keller et al., “Recurrence after recovery from major depressive disorder during 15 years of observational follow-up,” *American Journal of Psychiatry*, vol. 156, no. 7, pp. 1000–1006, 1999.
View at: Publisher Site | Google Scholar
14. L. L. Judd, P. J. Schettler, and H. S. Akiskal, “The prevalence, clinical relevance, and public health significance of subthreshold depressions,” *Psychiatric Clinics of North America*, vol. 25, no. 4, pp. 685–698, 2002.
View at: Publisher Site | Google Scholar
15. J. Angst and R. Sellaro, “Historical perspectives and natural history of bipolar disorder,” *Biological Psychiatry*, vol. 48, no. 6, pp. 445–457, 2000.
View at: Publisher Site | Google Scholar
16. M. Olfson, R. C. Kessler, P. A. Berglund, and E. Lin, “Psychiatric disorder onset and first treatment contact in the United States and ontario,” *American Journal of Psychiatry*, vol. 155, no. 10, pp. 1415–1422, 1998.
View at: Publisher Site | Google Scholar
17. M. J. Zaki, J. T. L. Wang, and H. T. T. Toivonen, “Data mining in bioinformatics,” *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 2, pp. 198-199, 2003.
View at: Publisher Site | Google Scholar
18. D. Chauhan and V. Jaiswal, “An efficient data mining classification approach for detecting lung cancer disease,” in *Proceedings of the 2016 International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, October 2016.
View at: Publisher Site | Google Scholar
19. R. Polikar, “Ensemble learning,” *Scholarpedia*, vol. 4, no. 1, p. 2776, 2009.
View at: Publisher Site | Google Scholar
20. D. Greenstein, J. D. Malley, B. Weisinger, L. Clasen, and N. Gogtay, “Using multivariate machine learning methods and structural MRI to classify childhood onset schizophrenia and healthy controls,” *Frontiers in Psychiatry*, vol. 3, 2012.
View at: Publisher Site | Google Scholar

21. Y. T. Jo, S. W. Joo, S. H. Shon, H. Kim, Y. Kim, and J. Lee, "Diagnosing schizophrenia with network analysis and a machine learning method," *International Journal of Methods in Psychiatric Research*, vol. 29, no. 1, 2020.
View at: Publisher Site | Google Scholar
22. H. Yang, J. Liu, J. Sui, G. Pearlson, and V. D. Calhoun, "A hybrid machine learning method for fusing fMRI and genetic data: combining both improves classification of schizophrenia," *Frontiers in Human Neuroscience*, vol. 4, 2010.
View at: Publisher Site | Google Scholar
23. S. Srinivasagopalan, J. Barry, V. Gurupur, and S. Thankachan, "A deep learning approach for diagnosing schizophrenic patients," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 31, no. 6, pp. 803–816, 2019.
View at: Publisher Site | Google Scholar
24. R. N. Pläschke, E. C. Cieslik, V. I. Müller et al., "On the integrity of functional brain networks in schizophrenia, Parkinson's disease, and advanced age: evidence from connectivity-based single-subject classification," *Human Brain Mapping*, vol. 38, no. 12, pp. 5845–5858, 2017.
View at: Publisher Site | Google Scholar
25. W. H. L. Pinaya, A. Gadelha, O. M. Doyle et al., "Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia," *Scientific Reports*, vol. 6, no. 1, 2016.
View at: Publisher Site | Google Scholar
26. W. H. L. Pinaya, A. Mechelli, and J. R. Sato, "Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: a large-scale multi-sample study," *Human Brain Mapping*, vol. 40, no. 3, pp. 944–954, 2018.
View at: Publisher Site | Google Scholar
27. A. M. Chekroud, R. J. Zotti, Z. Shehzad et al., "Cross-trial prediction of treatment outcome in depression: a machine learning approach," *The Lancet Psychiatry*, vol. 3, no. 3, pp. 243–250, 2016.
View at: Publisher Site | Google Scholar
28. A. Ahmed, R. Sultana, M. T. R. Ullas, M. Begom, M. M. I. Rahi, and M. A. Alam, "A machine learning approach to detect depression and anxiety using supervised learning," in *Proceedings of the 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, Gold Coast, Australia, 2019.
View at: Google Scholar
29. A. Sau and I. Bhakta, "Predicting anxiety and depression in elderly patients using machine learning technology," *Healthcare Technology Letters*, vol. 4, no. 6, pp. 238–243, 2017.
View at: Publisher Site | Google Scholar

30. C. D. Katsis, N. S. Katertsidis, and D. I. Fotiadis, "An integrated system based on physiological signals for the assessment of affective states in patients with anxiety disorders," *Biomedical Signal Processing and Control*, vol. 6, no. 3, pp. 261–268, 2011.
View at: Publisher Site | Google Scholar
31. A. Sau and I. Bhakta, "Screening of anxiety and depression among seafarers using machine learning technology," *Informatics in Medicine Unlocked*, vol. 16, Article ID 100228, 2019.
View at: Publisher Site | Google Scholar
32. K. Hilbert, U. Lueken, M. Muehlhan, and K. Beesdo-Baum, "Separating generalized anxiety disorder from major depression using clinical, hormonal, and structural MRI data: a multimodal machine learning study," *Brain and Behavior*, vol. 7, no. 3, Article ID e00633, 2017.
View at: Publisher Site | Google Scholar
33. A. J. Xu, M. A. Flannery, Y. Gao, and Y. Wu, "Machine learning for mental health detection," 2019, https://digitalcommons.wpi.edu/mqp-all/6732/.
View at: Google Scholar
34. V. Rocha-Rego, J. Jogia, A. F. Marquand, J. Mourao-Miranda, A. Simmons, and S. Frangou, "Examination of the predictive value of structural magnetic resonance scans in bipolar disorder: a pattern classification approach," *Psychological Medicine*, vol. 44, no. 3, pp. 519–532, 2013.
View at: Publisher Site | Google Scholar
35. D. Grotegerd, T. Suslow, J. Bauer et al., "Discriminating unipolar and bipolar depression by means of fMRI and pattern classification: a pilot study," *European Archives of Psychiatry and Clinical Neuroscience*, vol. 263, no. 2, pp. 119–131, 2012.
View at: Publisher Site | Google Scholar
36. G. Valenza, M. Nardelli, A. Lanata et al., "Predicting mood changes in bipolar disorder through heartbeat nonlinear dynamics," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 4, pp. 1034–1043, 2016.
View at: Publisher Site | Google Scholar
37. J. Mourão-Miranda, J. R. Almeida, S. Hassel et al., "Pattern recognition analyses of brain activation elicited by happy and neutral faces in unipolar and bipolar depression," *Bipolar Disorders*, vol. 14, no. 4, pp. 451–460, 2012.
View at: Publisher Site | Google Scholar
38. G. Roberts, A. Lord, A. Frankland et al., "Functional dysconnection of the inferior frontal gyrus in young people with bipolar disorder or at genetic high risk," *Biological Psychiatry*, vol. 81, no. 8, pp. 718–727, 2017.
View at: Publisher Site | Google Scholar

39. G. Akinci, E. Polat, and O. M. Koçak, "A video based eye detection system for bipolar disorder diagnosis," in *Proceedings of the 2012 20th Signal Processing and Communications Applications Conference (SIU)*, Mugla, Turkey, April 2012.
View at: Publisher Site | Google Scholar
40. M.-J. Wu, I. C. Passos, I. E. Bauer et al., "Individualized identification of euthymic bipolar disorder using the cambridge neuropsychological test automated battery (CANTAB) and machine learning," *Journal of Affective Disorders*, vol. 192, pp. 219–225, 2016.
View at: Publisher Site | Google Scholar
41. 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, 2017.
View at: Publisher Site | Google Scholar
42. D. Leightley, V. Williamson, J. Darby, and N. T. Fear, "Identifying probable post-traumatic stress disorder: applying supervised machine learning to data from a UK military cohort," *Journal of Mental Health*, vol. 28, no. 1, pp. 34–41, 2018.
View at: Publisher Site | Google Scholar
43. S. Papini, D. Pisner, J. Shumake et al., "Ensemble machine learning prediction of posttraumatic stress disorder screening status after emergency room hospitalization," *Journal of Anxiety Disorders*, vol. 60, pp. 35–42, 2018.
View at: Publisher Site | Google Scholar
44. D. Conrad, S. Wilker, A. Pfeiffer et al., "Does trauma event type matter in the assessment of traumatic load?" *European Journal of Psychotraumatology*, vol. 8, no. 1, Article ID 1344079, 2017.
View at: Publisher Site | Google Scholar
45. C. R. Marmar, A. D. Brown, M. Qian et al., "Speech-based markers for posttraumatic stress disorder in US veterans," *Depression and Anxiety*, vol. 36, no. 7, pp. 607–616, 2019.
View at: Publisher Site | Google Scholar
46. D. Vergyri, B. Knoth, E. Shriberg et al., "Speech-based assessment of PTSD in a military population using diverse feature classes," in *Proceedings of the INTERSPEECH 2015*, Dresden, Germany, 2015.
View at: Publisher Site | Google Scholar
47. L. E. Salminen, R. A. Morey, B. C. Riedel, N. Jahanshad, E. L. Dennis, and P. M. Thompson, "Adaptive identification of cortical and subcortical imaging markers of early life stress and posttraumatic stress disorder," *Journal of Neuroimaging*, vol. 29, no. 3, pp. 335–343, 2019.
View at: Publisher Site | Google Scholar
48. D. Rangaprakash, G. Deshpande, T. A. Daniel et al., "Compromised hippocampus-striatum pathway as a potential imaging biomarker of mild-traumatic brain injury and

posttraumatic stress disorder," *Human Brain Mapping*, vol. 38, no. 6, pp. 2843–2864, 2017.
View at: Publisher Site | Google Scholar

49. M. Sumathi and B. Poorna, "Prediction of mental health problems among children using machine learning techniques," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 1, 2016.
View at: Publisher Site | Google Scholar
50. A. E. Tate, R. C. McCabe, H. Larsson, S. Lundström, P. Lichtenstein, and R. Kuja-Halkola, "Predicting mental health problems in adolescence using machine learning techniques," *PLoS One*, vol. 15, no. 4, Article ID e0230389, 2020.
View at: Publisher Site | Google Scholar