A Machine Learning Technique to Analyze Depressive Disorders

Dixita Mali
Techno India NJR Institute of Technology
Kritika Kumawat
Techno India NJR Institute of Technology
Gaurav Kumawat
Techno India NJR Institute of Technology
Prasun Chakrabarti
Techno India NJR Institute of Technology
Sandeep Poddar
Lincolns university College
Tulika Chakrabarti
Sir Padampat Singhania University
Jemal Hussaine
Deakin University
Ali-Mohammad Kamali
Shiraz University of Medical Sciences
Vadim Bolsev
Federal Scientific Agroengineering Center VIM: FGBNU Federal'nyj naucnyj agroinzenernyj centr VIM
Babak Kateb
Society for Brain Mapping and Therapeutics
Mohammad Nami (torabinami@sums.ac.ir)
Shiraz University of Medical Sciences https://orcid.org/0000-0003-1410-5340

Research Article

Keywords: Depression, XGBoost Tree, Random Trees, Neural Network, Random Forest, C5.0

Posted Date: March 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-322564/v1

Abstract

Depression is an ordinary mental health care problem and the usual cause of disability worldwide. The main purpose of this research was to determine that how depression affects the life of an individual. It is a leading cause of morbidity and death. Over the last 50–60 years, large numbers of studies published various aspects including the impact of depression. The main purpose of this research is to determine whether the person is suffering from depression or not. The dataset of Depression has been taken from the Kaggle website. Guided Machine Learning classifiers have helped in the highest accuracy of a dataset. Classifiers like XGBoost Tree, Random Trees, Neural Network, SVM, Random Forest, C5.0, and Bay Net. From the result, it is evident that the C5.0 classifier is giving the highest accuracy with 83.94 % and for each classifier, the result is derived based without pre-processing.

1. Introduction

Depression has now become a common disease for the people nowadays. It is especially seen in youngsters due to several reasons. We feel moody, weakness, loss of energy, we can't able to take proper sleep, we also feel disturbed by society. We are unable to handle our responsibilities. It is a type of disease from which everyone is suffering it may be due to responsibilities, ignorance in life, and due to many other reasons.

A mood disorder can also be a symptom of depression, we can't able to feel too fresh, people become moody. Antidepressant medicines and psychotherapies become an effective treatment for depression. If this problem become continues for a long time then it leads to a great effect in relation and also leads to mental weakness. That’s why it's recommended to treat mental disorders as soon as possible.

There is various machine learning algorithm that is used for the prediction that person is suffering from depression or not. Our main aim is to determine the accuracy of various classifier algorithms of machine learning and find out the algorithm which is best fitted for our dataset. We selected the following classifier for finding the accuracy of the test data: - XGBoost Tree, Random Trees, Neural Network, SVM, Random Forest, C5.0, Bay Net, and Random Tree.

2. Issue Statement And Background Knowledge

According to facial and verbal analysis techniques presented the algorithm with the help of upgraded classification of the data. An average detection of 82.2% in males and 70.5% in females are recorded by the system [1].

Stolar *et al.* determined the advanced spectral roll-off set in improvement with the help of phonic spectral features. All the features that included the best individual spectral gave an average classification with the accuracy of 71.4% in males and 70.6% in females [2].

Four common classification prototypes, including Bayes Network, C 4.5 Decision Tree, Support Vector Machine (SVM), and Artificial Neural Network (ANN) were applied to determine the aging patients who were suffering from the prior symptoms of depression and found out the ANN showing the best results showed by Soundariya *et al.* [3].

To check the disclosure of depression Tsugawa *et al.* examined the activities of the user in social media. Through experiments, they showed features acquired from the activities of users which helped to anticipate depression of users with 69% accuracy[4].

Haque *et al.* explored the 3D facial features and the language vocalized to gauge the depression intensity. The embedded Convolutional Neural Network (CNN) model had been compared by this research. This model denoted a sensitivity of 83.3% and specificity of 82.6% [5].

Aldarwish and Ahmed applied Naive Bayes and SVM models on the prior processed posts from social sites. To classify SNS users they came up with a web application that could be used by depressed patients and psychiatrists. By training and formulating better models there are chances to increase the accuracy of this model in further modifications [6].

De Choudhury *et al.* evolved an estimated accuracy that could be acquired by using activities of the depressed users on Twitter. For Machine Learning they obtained the training data with the help of numerous people. By using SVM they recorded the activities of the users on Twitter to predict the risk of depression among them. Experimental results showed an approximation of 70% accuracy[7].

To detect the sign of depression in a person's tweet Shetty *et al.* employed classifier Machine Learning on the Twitter data set. From a developer's Twitter account, they obtained Twitter posts by using the Twitter API [8].

Cao *et al.* obtained an accuracy of 84.21%. They classified seriously depressed patients by feature selection and SVM model formulated on functional connections of resting-state FMRI [9].

Liao *et al.* applied SVM to classify severe depression patients which were centred on resting-state EEG signals, and hence obtained an accuracy of 80%[10].

3. Methodology

3.1 Dataset Introduction

This dataset analysis depression among people. This is now a common disease among people. Most people around the world are suffering from depression. This dataset analyses the depression point based on age, sex, the status of marriage, education, number of children, total members, etc.

The model is based on the following attribute: -

- Person's durable asset

- Person's save asset

Table 1 Attribute Description of Dataset

S.No.	Attribute Name	Description
1.	Age	Person's age (It is in years)
2.	Married	Persons' marital status (Yes/No)
3.	Number of Children's	Number of children's (It is in numbers)
4.	Education	Education Status
5.	Total Members	Total Members in the family (It is in numbers)
6.	Gained Asset	Person's gained asset (It is the amount of gained asset)
7.	Durable Asset	Person's durable asset (It is the amount of durable)
8.	Save Asset	Person's saved asset (It is the amount saved asset)
9.	Living Expenses	Person's living expenses (It is the amount of money)
10.	Other Expenses	Person's other expenses (It is the amount of money)
11.	Incoming Salary	Person's salary (Yes/No)
12.	Incoming Own Farm	A person has its own farm (Yes/No)
13.	Incoming Business	A person has its own business (Yes/No)
14.	Incoming No Business	Person's incoming_no_buisness(Yes/No)
15.	Incoming Agricultural	Person's agriculture
16.	Farm Expenses	Person's farm expenses (It is the amount of money)
17.	Labor Primary	A person has labor (Yes/No)
18.	Lasting Investment	Person's lasting investment (It is the amount of money)
19.	No Lasting Investment	Person's no lasting investment (It is the amount of money)
20.	Depressed	Target Variable (Yes/No)

3.2 Classification Algorithm

3.2.1 Random Tree

It is a super algorithm that is useful for both regression and classification. This algorithm creates multiple decision trees and get prediction by each of them and finally selects the best solution. It is less accurate than the XG boost tree. It is created by using the random subspace method. In this method, multiple deep trees are trained in different parts of the same data set to achieve less variance.

3.2.2 SVM

SVM algorithm creates a decision boundary that segregates n-dimensional space into classes to put the new data point incorrect category. The best day which it chooses is call hyperplane. I also choose the extreme points/vectors which are called the support vectors that's why this algorithm is called Support Vector Machine.

3.2.3 C5.0

It is a calculation used to create a decision tree based on Quinlan's previous ID3 calculation. It is much easier to understand and deploy.

3.2.4 Random Forest

This algorithm simply generates multiple decision trees and further divides them into the class prediction and all the forest trees give a vote and then finally majority decision tree is chosen by this algorithm and we get the final result. The method used by this algorithm is "bagging".

3.2.5 Bay Net

It is a probabilistic graphical model also known as a decision tree, Bayesian network classifier, and recognized by many other names. It depends on Bay's Theorem. It assumes that the presence/absence of any features of a variable is not related to the presence/absence of features of other variables.

3.2.6. XGBoost Tree

XGBoost means "Extreme Gradient Boosting". This algorithm uses a gradient boosting framework. It is used for supervised learning in Machine Learning. It performs well when the prediction involves unstructured data such as images and text.

3.2.7 Neural Network

This algorithm establishes a relationship within the dataset in the way the hum brain does, Neural system is similar to the system of neurons. It may be organic or artificial. It changes the input data to generate the best possible network so there is no need to redesign the output criteria.

3.3 Performance Evaluation Measure

3.3.1 Confusion Matrix

Classification Matrix describes the performance of a classification model in the tabular format on a set of data for which we know the true value.

3.3.2 Classification Accuracy

It is the rate of correct classification. It may for an independent test set or using some variation of the cross-validation idea.

3.3.3 Classification Error

Classification errors come when g(X) ≠ Y. The best classifier g*, known as the Bays classifier, and it is one that minimizes the probability of classification error.

3.3.4 Precision

It is the closeness of more than two measurements. If you get a nearby value like 3.2 each time then your result would be precise. Precision is not dependent on accuracy. You may be precise but inaccurate.

3.3.5 Recall

It the ratio of how many times you get the correct result to the number of results.

3.3.6 AUC

AUC means “Area Under the ROC Curve”. It measures all the 2-dimensional area under the ROC curve and measures the performance across all classification thresholds.

3.3.7 GINI

It is the probability of wrongly classified variables when randomly chosen.

4. Result And Discussion

There are two partitions of the dataset testing and training. IBM SPSS Modeler is used to find out the result and this dataset is 70% trained and 30% tested. 7 classifiers were used to find out the most accurate result. For each classifier, results are noted based on – (i) without SMOTE (ii) without SMOTE AUC (iii) Without SMOTE F-Measure (iv)Without SMOTE PRA.

The results of the models are as follows:

The following table consists of precision, recall, F-measure, AUC, GINI coefficient, and accuracy values: -

Table 2: Results after analysis of the above model

	Precision	Recall	F-Measure	AUC	GINI Coefficient	Accuracy
Random Tree	0.867	0.852	0.860	0.60	0.206	76.61%
SVM	0.843	0.900	0.871	0.57	0.131	76.38%
Neural Network	0.837	1.000	0.911	0.55	0.890	82.34%
Bay Net	0.845	0.952	0.895	0.55	0.103	**79.13%**
Random Forest	0.841	0.961	0.897	0.51	0.031	80.28%
XGBoost Tree	0.842	0.989	0.910	0.52	0.031	**83.49%**
C5.0	0.839	1.000	0.913	0.50	0.000	83.94%

According to the AUC values of all the classifiers, the graph can be represented as:

Table 3: AUC values of all the classifiers

	AUC
Random Forest	0.60
SVM	0.57
Neural Network	0.55
Bay Net	0.55
Random Forest	0.51
XGBoost Tree	0.52
C5.0	0.50

The following table consists of the precision, recall, and accuracy value of all classifier, based on these values a comparison graph can be represented as: -

Table 4:Precision, Recall, and Accuracy value of all classifier

	Precision	Recall	Accuracy
Random Tree	0.867	0.852	76.61%
SVM	0.843	0.900	76.38%
Neural Network	0.837	1.000	82.34%
Bay Net	0.845	0.952	**79.13%**
Random Forest	0.841	0.961	80.28%
XGBoost Tree	0.842	0.989	**83.49%**
C5.0	0.839	1.000	83.94%

According to the F-Measure values of all the classifiers, the following graph can be represented as: -

Table 5: F-Measure values of all the classifiers

	F-Measure
Random Tree	0.860
SVM	0.871
Neural Network	0.911
Bay Net	0.895
Random Forest	0.897
XGBoost Tree	0.910
C5.0	0.913

5. Conclusion

Depression now becomes a super disease among people around the globe. Around 75% of the people were suffering from depression remain untreated in developing countries [11]. This paper aims to predict whether a person is suffering from depression or not. To achieve the best result, 7 classifiers are used such as Random Tree, SVM, Neural Network, Bay Net, Random Forest, XGBoost Tree, C5.0. The result is noted without applying any filters.

6. Future Work

This type of study helps in the future to prevent depression. This data helps in spreading a serious effect of depression and spread health awareness among people. If this study continues it will give us a better understanding of depression and better treatment for the people who are suffering from this. In the future,

by collecting more data and information we will get more accurate results. These studies will save people's health, relations, and money in a large amount. We hope that people will understand that depression is just not a part of life but it plays a major role in ruining a person's health and relation.

Declarations

7. Acknowledgment

The authors are thankful to Techno India NJR Institute of Technology, Udaipur, Rajasthan, India for providing necessary resources and infrastructure and encouragement to conduct this project work and to Lincoln University College, Malaysia for academic support.

References

[1] Ramalingam D, Sharma V, Zar P. Study of Depression Analysis using Machine Learning Techniques. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075,Volume-8, Issue-7C2, May 2019

[2] Stolar MN, Lech M, Stolar SJ, Allen NB. Detection of adolescent depression from speech using optimised spectral roll-off parameters. Biomedical Journal. 2018;2:10.

[3] Soundariya R S, Nivaashini M, Tharsanee R M, Thangaraj P. Application of Various Machine Learning Techniques in Sentiment Analysis for Depression Detection. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-10S, August 2019.

[4] Tsugawa S, Kikuchi Y, Kishino F, Nakajima K, Itoh Y, Ohsaki H. Recognizing depression from twitter activity. In Proceedings of the 33rd annual ACM conference on human factors in computing systems 2015 Apr 18 (pp. 3187-3196).

[5] Haque A, Guo M, Miner AS, Fei-Fei L. Measuring depression symptom severity from spoken language and 3D facial expressions. arXiv preprint arXiv:1811.08592. 2018 Nov 21.

[6] Aldarwish MM, Ahmad HF. Predicting depression levels using social media posts. In2017 IEEE 13th international Symposium on Autonomous decentralized system (ISADS) 2017 Mar 22 (pp. 277-280). IEEE.

[7] De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. In Proceedings of the 7th International AAAI Conference on Social Media and Weblogs (ICWSM'13) (July 2013).

[8] Shetty NP, Muniyal B, Anand A, Kumar S, Prabhu S. Predicting depression using deep learning and ensemble algorithms on raw twitter data. International Journal of Electrical & Computer Engineering (2088-8708). 2020 Aug 1;10.

[9] Cao L, Guo S, Xue Z, Hu Y, Liu H, Mwansisya TE, Pu W, Yang B, Liu C, Feng J, Chen EY. Aberrant functional connectivity for diagnosis of major depressive disorder: a discriminant analysis. Psychiatry and clinical neurosciences. 2014 Feb;68(2):110-9.

[10] Liao SC, Wu CT, Huang HC, Cheng WT, Liu YH. Major depression detection from EEG signals using kernel eigen-filter-bank common spatial patterns. Sensors. 2017 Jun;17(6):1385.

[11] World Health Organization (WHO) (2017). Depression. Fact sheet. Available online at: http://www.who.int/mediacentre/factsheets/fs369/en/

Figures

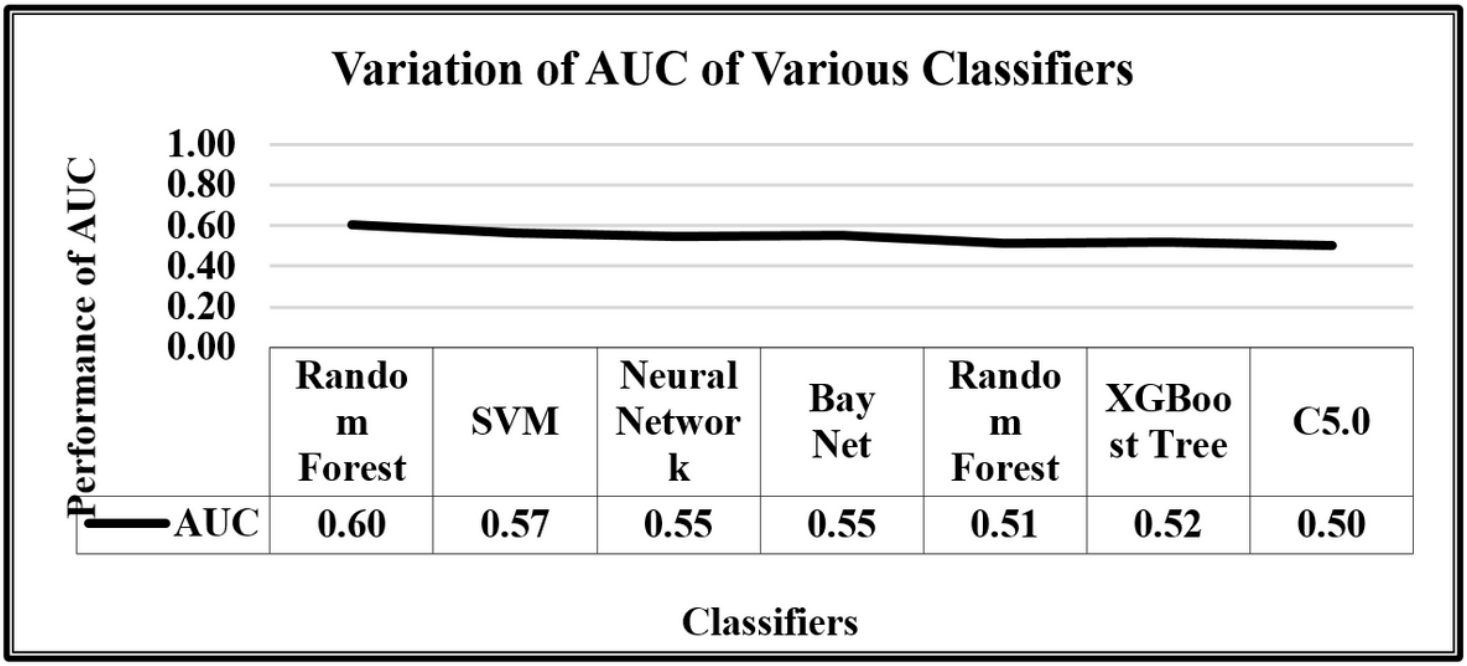

Figure 1

AUC grid

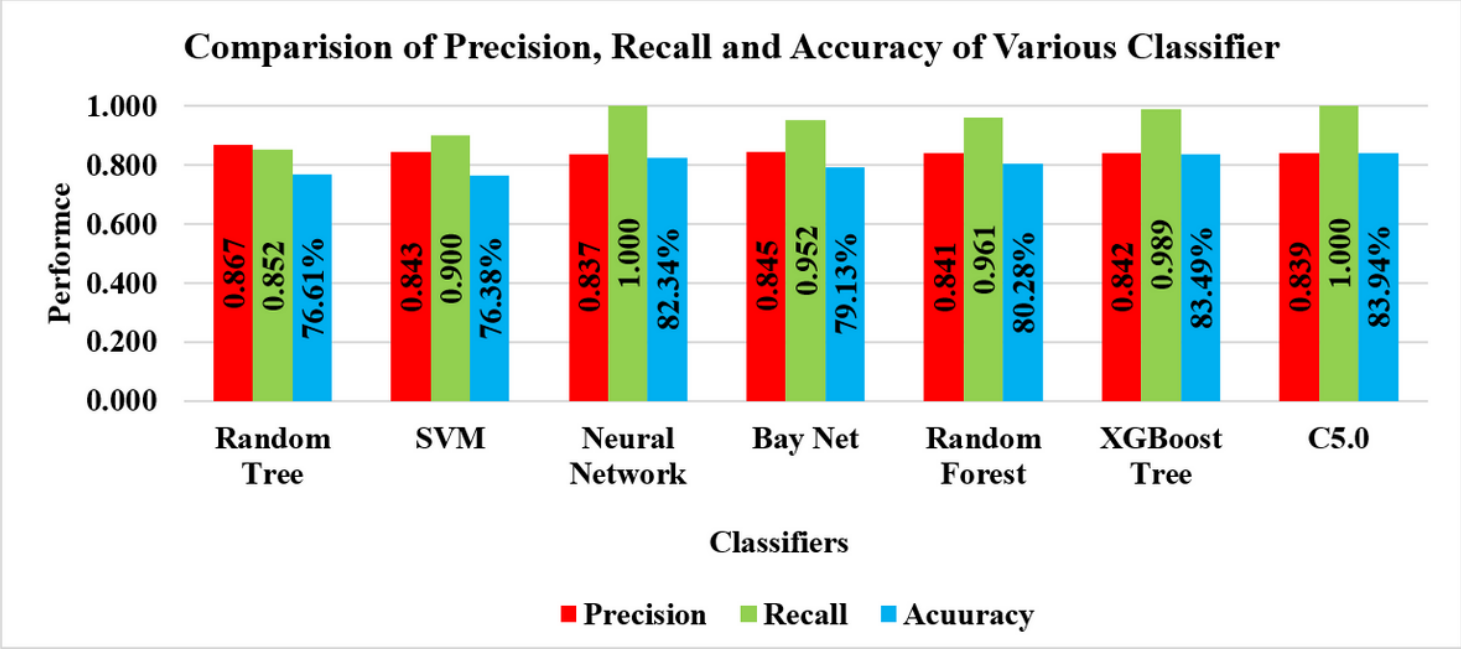

Figure 2

Precision, Recall and Accuracy Graph

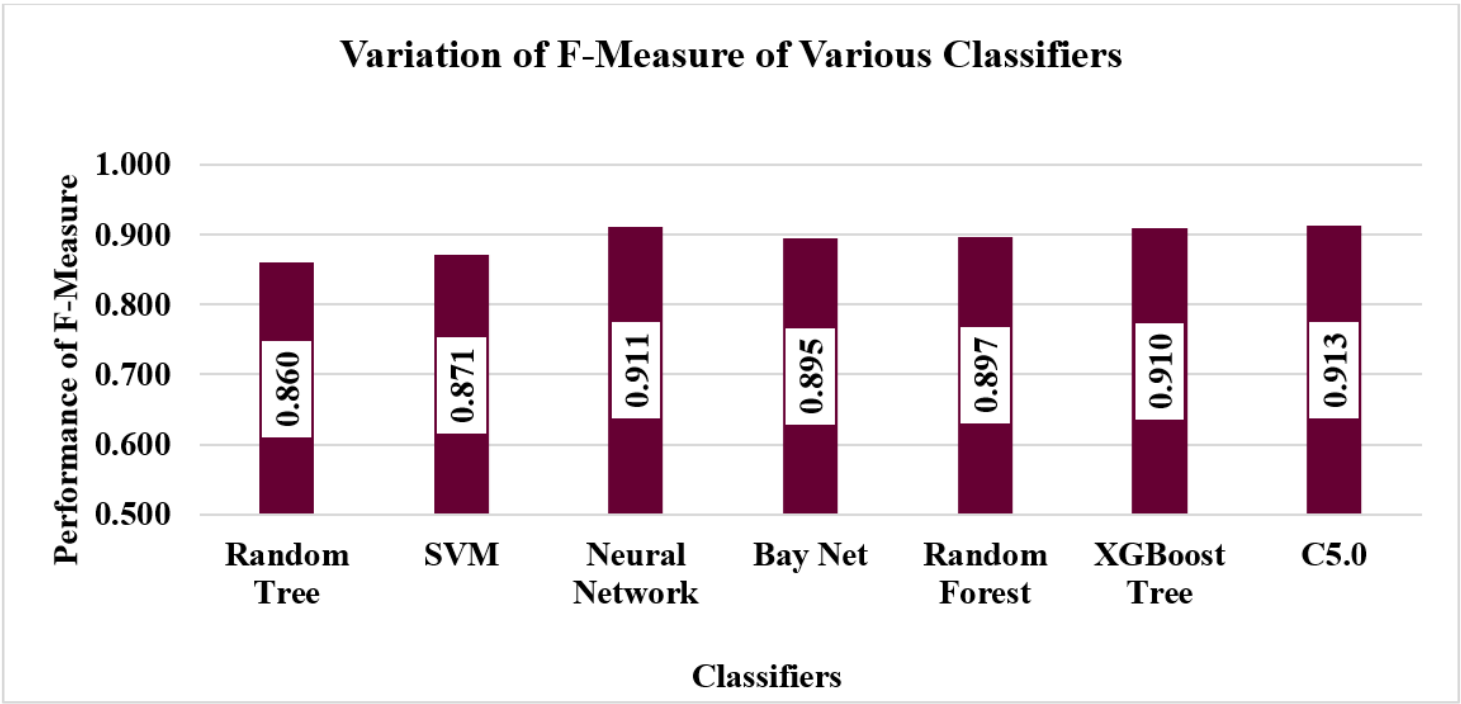

Figure 3

F- measure Grid