International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 08 | Aug 2023 www.irjet.net p-ISSN: 2395-0072

Mental Illness Prediction using Machine Learning Algorithms

Falguni Wani[1], Ved Deore[2], Shivam Gorane[3], Santosh Chobe[4]

[1,2,3]Student, Dept. of Computer Engineering, Pimpri Chinchwad College of Engineering and Research, Ravet, Pune, Maharashtra, India
[4]Professor, Dept. of Computer Engineering, Pimpri Chinchwad College of Engineering and Research, Ravet, Pune, Maharashtra, India

---

Abstract - *Depression is one of the most concerned issues in the society and it is not limited to certain age of a person. Depression management is an approach for analyzing and working on these concerns and lead to quality of life. The idea behind this work is to analyze depression, anxiety and stress based on some psychological test like Depression Anxiety Stress Scale-21(DASS 21). Machine learning is an emerging field in computer science and has ability to predict outcome based on certain situations or inputs. Machine learning algorithms are used to predict depression, anxiety and stress levels by using standard psychological scale. Training and testing datasets are used to train and test the developed machine learning model. Various machine learning algorithms like Support Vector Machine, Random Forest, Naïve Bayes, etc. are implemented and compared in order to evaluate the best among all. The accuracy of the best algorithm is boosted using the boosting technique of ensemble learning method and a user interface is used for self-evaluation. From the classification algorithms used SVM has surpassed the other machine learning algorithms and then it is boosted using AdaBoost giving highest accuracy for prediction.*

Key Words: Depression, Anxiety, Stress, Classification, Depression Anxiety Stress Scale-21(DASS-21), Supervised Learning, AdaBoost, Support Vector Machine

1. INTRODUCTION

Healthcare is among the serious issues in front of the entire world regardless of any circumstances. As a ruling interest globally, besuited, well organized, effective and robust wellness systems are built to improve and conserve the quality standards of life. Anxiety, depression, stress, irritation and disappointment have become so normal that individuals now imagine them to be part of personal and professional life.

The World Health Organization (WHO) has estimated that 3.8% of the population experience depression, including 5% of adults (4% among men and 6% among women), and 5.7% of adults older than 60 years. Approximately 280 million people in the world have depression [1]. Differentiating between anxiety and depression is complicated for machines; therefore, a suitable machine learning algorithm is necessary for an applicable recognition.

Mental health is an integral and essential component of health. The WHO constitution states: "Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity."[2]. The leading symptoms of depression from a medical point of view are lack of concentration, loss of memory, loss of interest in recreational activities, an inability to make decisions, overeating and weight gain, weight loss, low appetite and irritation, etc. These symptoms have a significant effect on crucial areas of an individual's life.

The symptoms of anxiety are irritability, insomnia, nervousness, sweating, fatigue, panic, increased heart rate and a sense that something is about to happen, difficulty in concentrating and rapid breathing.

The common symptoms of stress are low energy levels, feeling upset or agitated, chronic headaches, impotence to relax, recurring overreaction and persistent colds or infections. Thus, anxiety, stress and depression have many common symptoms including fatigue, chest pain, insomnia, inability to concentrate and increased heart rate all of which makes classification tough for machines.

This paper is structured as follows: Section 2 explores related studies on anxiety, depression and stress along with the methods and techniques that were adopted. Section 3 describes the dataset used in the research herein, while Section 4 discusses the various classification algorithms. Section 5 studies the research gap found. Section 6 includes experimental setup used to perform this study, while section 7 describes the proposed system. Section 8 compares the results of machine learning algorithms. Finally, section 9 is the conclusion, which summarizes the study in its entirety.

2. LITERATURE REVIEW

The literature survey shows the study of various machine learning algorithms to predict depression, anxiety and stress.

In [3], Anu Priya, et. al. proposed machine learning model for predicting different levels of depression, anxiety and stress. They applied different machine learning algorithms like Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB), Random Forest (RF) and K-Nearest Neighbors (KNN). They also calculated different

comparison factors for choosing the best algorithm and found out that Naïve Bayes algorithm gives the best results. They used the Depression, Anxiety and Stress Scale questionnaire (DASS-21) to train and test their model. The size of dataset was less and there were imbalanced classes in the confusion matrix, so the model did not give the expected accuracy and the decision was made on the f1-score score so Random Forest was chosen to be the best fit for all three classes.

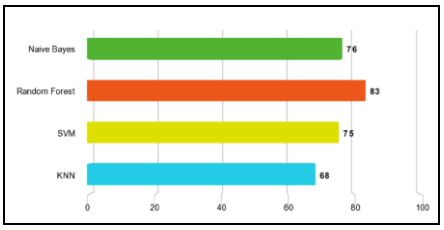

Fig -1: Accuracy chart for [3]

Astha Singh et. al. [4] proposed a model for identification of anxiety and depression. They collected the data by using standard DASS-21 questionnaire. They used some of the standard ML algorithms like decision tree (DT), SVM, Naïve Bayes, Random Forest along with KNN for training and testing purpose. Although the accuracy was not more than 95%, they selected SVM classifier as the best among all other. They also faced some problems related to dataset and affected its accuracy.

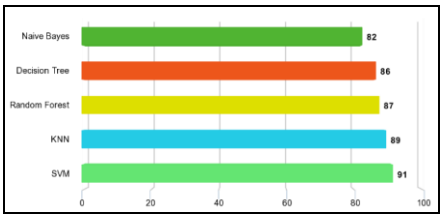

Fig -2: Accuracy Chart for [4]

Hritik Nandanwar et. al. [5] designed a model for depression prediction. The dataset used by them was collection of tweets from Twitter. They compared the performance of different machine learning models with labelled Twitter dataset. Different evaluation metrics like f1-score, recall and precision have been used to compare the performance. They got better results using Bag of Words with AdaBoost classifier.

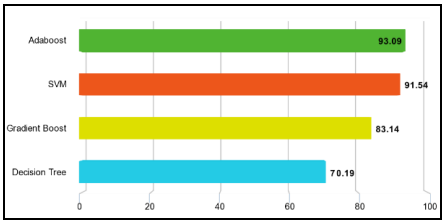

Fig -3: Accuracy Chart for [5]

Ruihu Wang [6] surveyed AdaBoost classifier for classification, feature selection, and their relation with support vector machine. They studied the fundamentals about AdaBoost algorithm for feature extraction and selection. They found out that AdaBoost algorithm gives better results and it has been widely used in many real time applications. The study also showed that one of the optimization algorithms known as Particle Swarm Optimization (PSO) also gives good results for prediction.

S Samanvitha, et. al. [7] built the model for depression detection using text data. They took data from different social media for building the model. As it is seen that people often express their feelings on online platforms. They tested their model with algorithms like Logistic Regression, Naïve Bayes, Random Forest and SVM classifier. They concluded that Naïve Bayes classifier gives the best results.

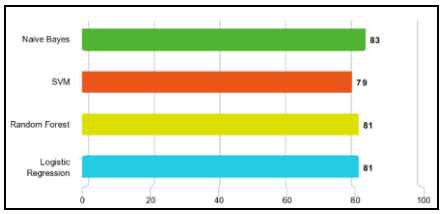

Fig -4: Accuracy Chart for [7]

Paphaychit Bounkeomany [8] designed a system for depression detection using speech. They used Adaboost-ELM framework which uses random numbers as they are number of meta-samples of dictionary atoms. They also enhanced this model using random dynamic integrated weighted classification model.

Ananna Saha et. al. [9] proposed a machine learning model of sentiment analysis of depressed person. They took the dataset of user generated contents from different social media applications like Facebook, Twitter. They have used python textblob package for different sentiment levels. They implemented Random Forest, Naïve Bayes classifier, DT, Sequential Minimal Optimization, Support Vector Machine (SVM), boosting technique such as Adaboost, Logistic Regression, Bagging, Multilayer Perceptron and Stacking algorithms. Among all they got 60.54% with Random Forest Algorithm.

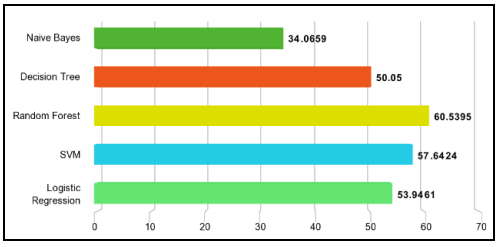

Fig -5: Accuracy Chart for [9]

Heidi Mochari-Greenberger et. al. [10] compared different psychological scales. That include DASS-21, GAD-7, PHQ-8. Their result of study shows that how GAD-7 and DASS-21 scale categorize severity of symptoms with respect to each other in population with behavioral and medical conditions of health. In this study, the authors concluded that DASS-21 remains the best scale to use as it consists of least number of questions to be answered.

Anju Prabha et. al. [11] used the Stroop Test mechanism to identify depression among people in COVID-19 situation. They found that there was a visible difference between the mental conditions of the patients between a certain time stimuli. They used machine learning algorithms such as Support Vector Machine (SVM), XGBoost, etc. to calculate the accuracy of the data and found that XGBoost algorithm gave the highest accuracy for the dataset. The XGBoost algorithm gave accuracy of 85.71% as compared to other algorithms used.

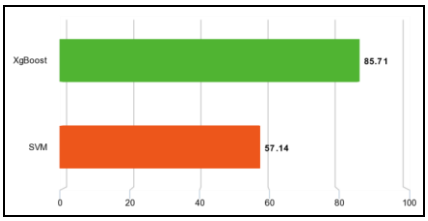

Fig -6: Accuracy chart for [11]

Shivangi Yadav et. al. [12] used Machine Learning and employed a wide range of Machine Learning algorithms to predict depression in people. They collected data by questioning people about their home, workplace environment and family history, etc. They used Machine Learning algorithms such as: K-Nearest Neighbors (KNN), Decision Tree, Multinomial Logistic Regression, Random Forest Classifier, Bagging, Boosting and Stacking. The best performance statistics was shown by boosting algorithm which gave accuracy of 81.75% which was then followed by Random Forest Classifier with accuracy of 81.22%.

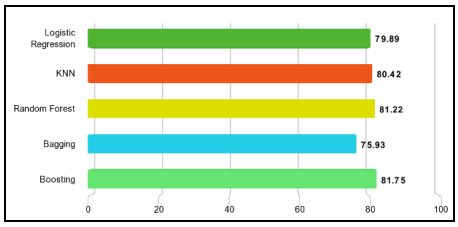

Fig -7: Accuracy chart for [12]

Md. Mehedi Hassan et. al. [13] have developed prediction models by classifying the dataset related to depression which were taken from Kaggle. They had primarily focused on feature selection. They selected the features after preprocessing the data and applied Logistic Regression, Correlation Matrix and Decision Tree methods. They applied different Machine Learning algorithms such as Logistic Regression, K-NN, SVM, and Naïve Bayes for building and classifying models. They got the best classification and accuracy for K-NN which is 79%. The other algorithms such as Logistic Regression, SVM, and Naive Bayes showed an accuracy of 77%, but K-NN was selected to be the best fit.

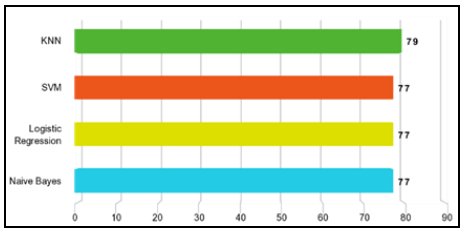

Fig -8: Accuracy chart for [13]

G H Suhas, et. al. [14] identified the risk of depression among people in the form of text. They collected and analyzed sentences from people to predict or detect whether the person is suffering through depression or not. They used different Machine Learning algorithms and found that Random Forest classifier gives the best accuracy when compared to CNN. Their system takes input from people and predict based on the responses.

Aanchal Bisht et. al. [15] proposed a methodology that will help teachers and parents to predict the levels of stress which students experience. They surveyed school children with a variety of 26 questions to analyze their stress levels and cure those using Machine Learning algorithms. They used different Machine Learning algorithms such as Decision Tree, Logistic Regression, K Nearest Neighbors and Random Forest. They found that K Nearest Neighbors (KNN) gave accuracy of 88% and proved to be the best for the implementation.

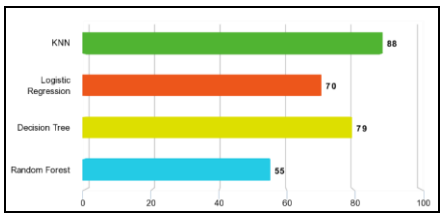

Fig -9: Accuracy chart for [15]

Akshada Kene et. al. [16] presented a study on previous research on stress detection using Machine Learning algorithms. They used the PhysioBank dataset to analyze different stress levels. They used statistical analysis for feature selection and extraction and found that gradient boost algorithm to be successful on the dataset used. The results demonstrated that the model displayed the accuracy of 83.33%, specificity of 75%, Sensitivity of 75%, Positive Recall value of 90%, and many more. The machine learning algorithms such as KNN, Random Forest, SVM and Naïve Bayes were used. Authors claimed Naïve Bayes model to be effective and efficient for stress classification and prediction with accuracy of 88%.

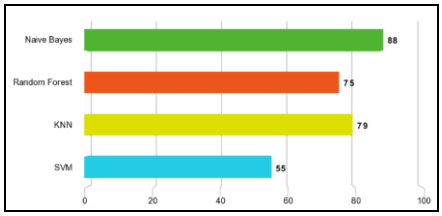

Fig -10: Accuracy chart for [16]

Somnath Sinha et. al. [17] have implemented Stress Prediction for the students and staff in the university premises to check whether they are stressed or stress-free. They used Machine Learning algorithms such as K Nearest Neighbors and Naïve Bayes to predict the results. They compared both the algorithms and tested them with easy manner. They concluded that Naïve Bayes algorithm is more efficient than KNN and has a high efficacy rate. The authors claimed the accuracy level over 94 percent for Naïve Bayes whereas the accuracy level for KNN as 87 percentage.

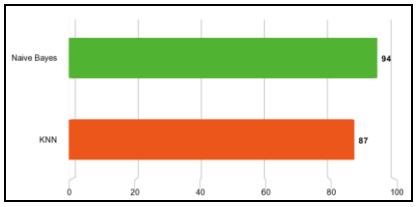

Fig -11: Accuracy chart for [17]

Anika Kapoor et. al. [18] in their research aimed to identify the anxiety disorders using Machine Learning Techniques. They identified symptoms such as Generalized Anxiety Disorder (GAD), Panic Disorder (PD), Post-Traumatic Stress Disorder, etc. They collected the dataset from many organizations/ institutions/ hospitals, etc. mainly through surveys and questionnaire related to the disease. For prediction they used Machine Learning algorithms such as Random Forest, Linear Regression, Support Vector Machine and others. Finally, they concluded that SVM has the highest accuracy for Generalised Anxiety Disorder (GAD) and Social Anxiety Disorder (SAD) while it was left behind by a margin of just 0.4% by GB Decision Tree (DT) for Post-Traumatic Stress Disorder (PTSD) and by 1% for Obsessive-Compulsive Disorder (OCD) by Random Forest (RF) which also achieved the best accuracy for Panic Disorder (PD).

Ahnaf Atef Choudhury et. al. [19] in their approach proposed predicting depression in university undergraduates and recommend them to the psychiatrist. They collected data from the after consultation with counselors, professors and psychologists. The authors found Random Forest to be the best algorithm followed by the Support Vector Machine (SVM) with accuracy around 73% and KNN with accuarcy 60% respectively. The Random Forest algorithm gave a better precision, recall and low false negatives. This research aimed to predict depression in early stages and ensure quick recovery for the victims to avoid any further mishaps.

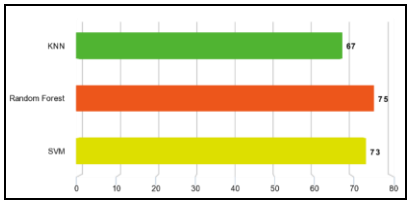

Fig -12: Accuracy chart for [19]

3. DATASET

The dataset consists of 21 questions based on the DASS-21[20] questionnaire, under the categories like, Depression, Anxiety, and Stress Scale. These questions are divided into the set of 7 for each category and the answer for each question is represented as numeric text as follows:

0 – Does not apply to me.

1 – Apply to me to some degree or sometimes.

2 – Apply to me to a considerable degree.

3 – Apply to me most of the time.

Table 1 shows the questions asked under each category.

Table -1: DASS-21 Questionnaire

Depression	Anxiety	Stress
I couldn't seem to experience any positive feeling at all.	I was aware of dryness of my mouth	I found it hard to wind down (calm down)
I found it difficult to work up the initiative to do things	I experienced breathing difficulty (e.g., excessively rapid breathing, breathlessness in the absence of physical exertion)	I tended to over-react to situations
I felt that I had nothing to look forward	I experienced trembling (shaking, e.g., in the hands)	I felt that I was using a lot of nervous energy (an excess of energy that you have when you are worried)
I felt down-hearted and blue (feeling sad and discouraged)	I was worried about situations in which I might panic and make a fool of myself	I found myself getting agitated (upset, disturbed)
I was unable to become enthusiastic	I felt I was close to panic	I found it difficult to relax

about anything		
I felt I was not worth much as a person	I was aware of the action of my heart in the absence of physical exertion (e.g., sense of heart rate increase, heart missing a beat)	I was intolerant of anything that kept me from getting on with what I was doing
I felt that life was meaningless	I felt scared without any good reason	I felt that I was rather touchy(sensitive)

4. CLASSIFICATION

4.1 Decision Tree

Decision Tree is Supervised Machine Learning algorithm which is used for classification as well as regression problems, but mostly it is used for classification problems. It is a tree structured classifier, where the features of the dataset are represented as internal nodes, decision rules are represented by the branches and outcome is represented by each leaf node.

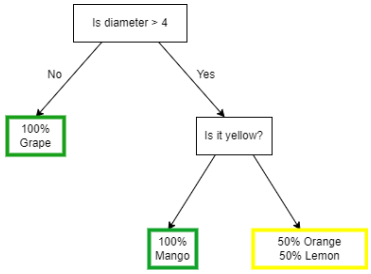

Fig -13: Decision Tree Example

4.2 Gaussian Naïve Bayes

Naïve Bayes classifier is a Supervised Machine Learning algorithm based on Bayes theorem and is used for solving classification problems. It includes training high dimensional dataset and is one of the simple and effective classification algorithms. It helps in building machine learning models that can make quick predictions. It is a probabilistic classifier and predicts results based on probability. The formula of Bayes theorem is as follows:

$$P(X|Y) = P(Y|X).\ P(X) / P(Y)$$

4.3 Random Forest

Random Forest is a Supervised Machine Learning algorithm which is used for Classification as well as Regression problems. It is a process of combining several classifiers to solve complex problems to improve the performance of the model. It contains a number of decision trees on various subsets of dataset and take the average to improve its accuracy. It can also handle dataset that contains continuous variables in regression and categorical variable in case of classification.

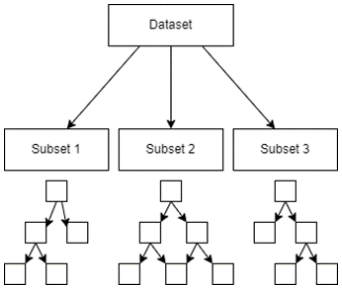

Fig -14: Random Forest Example

4.4 Support Vector Machine

Support Vector Machine (SVM) is a Supervised Machine Learning algorithm which is used for Classification as well as Regression problems. It is mostly used as Classification problems. It creates a decision boundary that can segregate n-dimensional spaces classes so that we can easily classify the data point in its correct category in future. SVM chooses extreme points to create the hyper plane and these extreme cases is termed as Support Vector Machine.

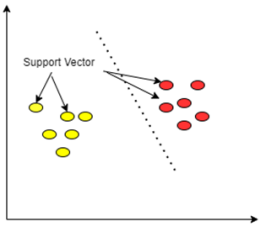

Fig -15: Support Vector Machine

4.5 XGBoost

XGBoost is an ensemble learning method that combines multiple weak classifiers into a stronger prediction model. XGBoost is also known as "Extreme Gradient Boosting". It also supports for parallel processing and one of its key features is its efficiency of handling missing values.

4.6 AdaBoost

AdaBoost is short form for "Adaptive Boosting". It is an ensemble learning technique which is used to make strong classifier based on the weak classifiers. It was first developed for the purpose of binary classification. The common estimator used with AdaBoost is decision tree with one level, i.e., decision tree with 1 split. These are also called as decision stumps.

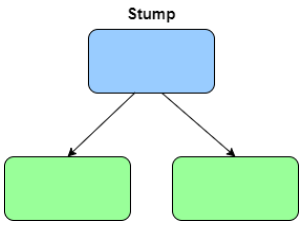

Fig -16: Adaptive Boosting

5. RESEARCH GAP

While studying the topic it was found that there are no boosting algorithms used with the DASS-21 scale. The accuracy given by various algorithms can be boosted using Adaboost or XGBoost algorithms. The boosted accuracy will help classify the problems more accurately.

6. EXPERIMENTAL SETUP

The dataset used is based on DASS21 standard questionnaire and the other supervised learning algorithms in this project. The system specifications are as follows:

- Processor: Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz 1.80 GHz
- RAM: 8GB
- System type: 64-bit Operating System

7. PROPOSED SYSTEM

Figure 17 depicts the proposed system. In the proposed system, the data is collected from the users/patients. This data is based on answers provided by the user/patient as per the standard questionnaire of Depression, Stress and Anxiety. The data has the features of the standard psychological factors. Next, the data pre-processing is done through handling missing values, transformation, encoding, etc. In the process of feature extraction, the strong and independent features are selected to achieve the target variable. Next, the model training is performed on the Machine Learning Algorithms such as: Random Forest, Support Vector Machine (SVM), Naïve Bayes, Decision Tree and XGBoost. It was found that SVM outperformed other algorithms. Then AdaBoost algorithm is used to boost the accuracy of the model. After applying the algorithms, the severity levels of the depression, stress and anxiety are calculated. Some tips on how to overcome the depression will be provided to the user/patient or some counselling may be provided. The performance of the model was measured with performance metrics viz. accuracy, recall, precision, f1-score and it was observed that the proposed system gives better results.

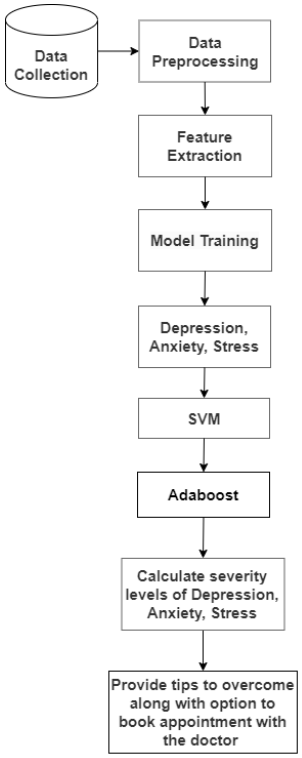

Fig -17: Proposed System Architecture

8. RESULT

Table 2 depicts the accuracy for different algorithms used for predicting severity levels of depression, anxiety and stress. Proposed system yielded the highest accuracy among all.

Table -2: Comparison Table

Algorithm	Accuracy		
	Depression	Anxiety	Stress
Naïve Bayes	75.81	56.38	72.54
Decision Tree	61.21	81.04	63.08
Random Forest	65.83	65.21	61.32
XGBoost	79.08	77.12	69.93
Support Vector Machine	86.27	89.54	89.54
Proposed System	94.08	92.89	93.49

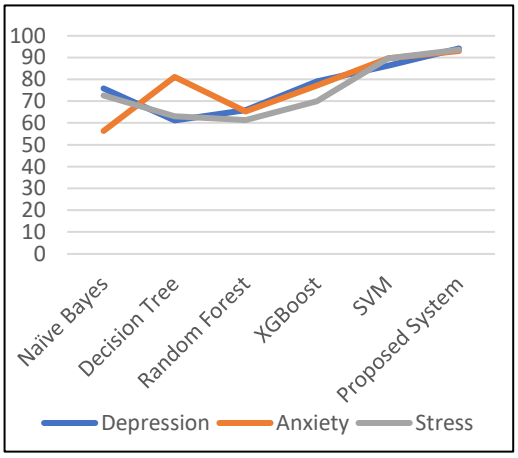

Chart -1: Accuracy Comparison Chart

9. CONCLUSION

From the study, it has been analysed that there are five levels of severity viz. Normal, Mild, Moderate, Severe and Extremely Severe for depression, stress and anxiety. The datasets used by various researchers were collected using a standard questionnaire to measure the frequent symptoms. The earlier research has shown an accuracy with single algorithm to a satisfactory level. The proposed system bestowed the accuracy of 94.08, 92.89, and 93.49 for Depression, Anxiety and Stress respectively.

10. REFERENCES

[1] https://www.who.int/news-room/fact-sheets/detail/depression", 01/04/2023

[2] https://www.who.int/data/gho/data/major-themes/health-and- well-being", 01/04/2023

[3] Anu Priya, Shruti Garg, Neha Prerna Tigga, "Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms" in International Conference on Computational Intelligence and Data Science (ICCIDS 2019), 2020

[4] Astha Singh, Divya Kumar, "Identification of Anxiety and Depression Using DASS-21 Questionnaire and Machine Learning" in First International Conference on Advances in Computing and Future Communication Technologies (ICACFCT), July 2022.

[5] Hritik Nandanwar, Sahiti Nallamolu, "Depression Prediction on Twitter using Machine Learning Algorithms" in 2nd Global Conference for Advancement in Technology (GCAT), November 2021.

[6] Ruihu Wang, "AdaBoost for Feature Selection, Classification and Its Relation with SVM, A Review" in International Conference on Solid State Devices and Materials Science, 2012.

[7] S Samanvitha; A R Bindiya, Shreya Sudhanva; B S Mahanand, "Naïve Bayes Classifier for depression detection using text data" in 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), February 2022.

[8] Paphaychit Bounkeomany, "Speech Major Depression Detection Based on Adaboost-ELM Algorithm" in IEEE International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), December 2020.

[9] Ananna Saha, Ahmed Al Marouf, Rafayet Hossain, "Sentiment Analysis from Depression-Related User-Generated Contents from social media" in 8th International Conference on Computer and Communication Engineering (ICCCE), July 2021.

[10] Heidi Mochari Greenberger, Reena L Pande, Aimee Peters, Lila Peters, Evie Andreopoulos, Naomi Pollock, "Comparison of DASS-21, PHQ-8, and GAD-7 in a virtual behavioral health care setting".

[11] Anju Prabha, Jyoti Yadav, Asha Rani, Vijander Signh, "A Pilot study for Depression Detection during COVID-19 using Stroop Test" in 8th International Conference on Signal Processing and Integrated Networks (SPIN), October 2021.

[12] Shivangi Yadav, Tanishk Kaim, Shobhit Gupta, "Predicting Depression from Routine Survey Data using Machine Learning" in 2nd International Conferences on Advances in Computing, Communication Control and Networking, March 2021.

[13] Md. Mehedi Hassan, Md. Asif Rakib Khan, Khan Kamrul Islam, Md. Mahedi Hassan, M M Fazle Rabbi, "Depression Detection System with Statistical Analysis and Data Mining Approaches", in 2021 International Conference on Science and Contemporary technologies, December 2021.

[14] G H Suhas, L Suraj, J Varun, D V Veda, H S Jayanna, "Machine Learning Approaches for Detecting Early Stage Depression using Text", in 2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques, February 2022.

[15] Aanchal Bhist, Shreya Vashisht, Muskan Gupta, Ena Jain, "Stress Prediction in Indian School Students using Machine Learning", in 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), August 2022.

[16] Akshada Kene, Shubhada Thakare, "Mental Stress Level Prediction and Classification based on Machine

Learning", in 2021 Smart Technologies, Communication and Robotics (STCR), November 2021.

[17] Somnath Sinha, Sriram R, "An Educational based Intelligent Student Stress Prediction using ML", in 2022 3rd International Conference for Emerging Technology (INCET), July 2022.

[18] Anika Kapoor, Shivani Goel, "Prediction of Anxiety Disorders using Machine Learning Techniques", in 2022 IEEE Bombay Section Signature Conference (IBSSC), February 2023.

[19] Ahnaf Atef Choudhury, Md. Rezwan Hassan Khan, Nabuat Zaman Nahim, Sadid Rafsun Talon, Samiul Islam, Amitabha Chakrabarty, "Predicting Depression in Bangladeshi Undergraduates using Machine Learning", in 2019 IEEE Region 10 Symposium (TENSYMP), January 2020.

[20] DASS-21 Lovibond PF, Lovibond SH. Manual for the Depression Anxiety & Stress Scales. 2nd ed. Sydney, Australia: Psychology Foundation; 1995.