DEPRESSION DETECTION USING SOCIO-DEMOGRAPHIC FEATURES AND PSYCHOSOCIAL FACTORS

Thasneem Vazim

Abstract

Depression, often known as major depressive disorder (MDD), is one of the most common psychiatric disorders worldwide. According to the WHO report, 322 million individuals worldwide suffer from depression, accounting for 4.4 percent of the global population. But unfortunately large percentage of cases of depression go undiagnosed and, thus, untreated. The data for this research was collected by a survey conducted among the Keralites designed after consultation with psychologists and professors. We used Natural Language Processing (NLP) techniques and machine learning methodologies to train the data and evaluate the efficiency of the proposed model. This study looked at six different machine learning classifiers that used a variety of socio-demographic and psychological data to determine whether or not a person is depressed. The Synthetic Minority Oversampling Technique (SMOTE), which minimises the class imbalance of the training data, was utilised to improve accuracy in predicting depression. The feature extraction techniques are used for extracting the most important socio-demographic and psychosocial factors responsible for forming depression. the GradientBoost classifier with the MRMR feature selection technique was found to be the the perfect model to predict depression among the participants. It has obtained an accuracy of 80

Keywords: Machine learning ,Depression,SMOTE,Feature Selection.

1 Introduction

Depression is the leading cause of global disability and is also a leading cause of suicide. Despite this, many people who suffer from depression are not treated for a variety of reasons. Timely identification of depressed symptoms, followed by assessment and therapy, can greatly improve the chances of controlling symptoms and the underlying condition, as well as attenuate harmful effects on personal, economic, and social life.

In psychological analysis and psychometrics, statistical inferences have been employed for decades. The use of Machine Learning (ML) in psychometrics has attracted media attention after the Cambridge Analytica affair[6]. Machine learning is a collection of algorithms that can automatically recognise patterns in data and utilise those patterns to forecast future data. Least-squares regression techniques are extensively used by psychologists to find patterns in data. ML analysis of experimental data, in contrast to statistical inference, is model agnostic and largely focuses on prediction rather than inference[6]. Due to the replicability problems of statistical inference, researchers are tending to machine learning from statistical inferences in psychometrics and psychological analysis nowadays. However, when compared to other fields, ML is not often employed in the analysis of psychological trials (e.g., genetics)[4][5]. Moreover, machine learning can help with time-consuming and complex jobs in healthcare field. These innovative technologies help to save money and improve therapeutic outcomes by accelerating proper medication discovery.

The development of depression in people is influenced by a variety of socio-demographic factors as well as psychosocial features. Age, sex, marital status, socioeconomic conditions, family environment, literacy, job security, depression history, and chronic medical conditions are all strongly linked to depression [8-10]. These variables can be used to construct an automated depression prediction system.

With the rapid growth of information and technology, the use of machine learning algorithms to infer meaningful patterns from data from various sectors is becoming increasingly important. Although machine learning algorithms are commonly employed in the medical and health fields, they are rarely applied in the psychological field. This research tries to establish the level of depression a person possess, as well as the major elements that cause depression and the best machine learning approach for detecting depressed people.

2 Related works

Tadesse MM et al. (2019) used various text classification algorithms to investigate the performance of both single feature and combination feature sets to measure the indications of depression. The findings suggest that good feature selection and diverse feature combinations lead to better prediction performance. The MLP classifier achieved 91 percent accuracy and the 0.93 F1 score achieved the greatest performance degree for identifying the existence of depression in Reddit social media in our investigation, demonstrating the strength and usefulness of integrated characteristics[1].

Priya et al. (2020) used machine learning algorithms to collect data on anxiety, stress, and depression in order to diagnose these mental diseases. They discovered that the Naive Bayes classifier performed best in predicting depression, with an accuracy of 85.50 percent[2].

Prince Kumar et al.(2020) used eight distinct machine learning models to predict five severity levels of anxiety, depression, and stress using the online DASS42 tool. These approaches are divided into four categories: Bayes, neural networks, lazy methods, and tree methods. The final methodology is a mix of the K-star and random forest methods. The hybrid strategy enhanced single algorithm accuracy, but it took 30 to 45 minutes to complete, whereas single algorithms required no more than five minutes.

Sau et al. (2017) focused on the geriatric population in their research. They used ten different machine learning classifiers to predict the onset of depression in older people. For the classification, the socio-demographic and health-related characteristics of geriatric patients were gathered. The Random Forest performed the best out of the ten classifiers.[22]

Hatton et al. (2019) used 284 senior individuals' psychometric and demographic data to predict the prevalence of depression. They evaluated the performance of the Extreme Gradient Boosting method with that of the Logistic Regression model in predicting the persistence of depression. Extreme Gradient Boosting outperformed Logistic Regression, according to the researchers.[3]

3 Methodology

3.1 Data acquisition

A survey was conducted for collecting the data from Keralites of different age groups. A questionnaire consisting of 45 questions were designed for the analysis. The first 30 questions were designed for collecting psychosocial, and socio-demographic data of the participants, and the last 15 questions were taken from the Burns Depression Checklist (BDC). The survey was conducted in the period between April 2022 and May 2022. The dataset consists of the responses of individuals.

This study used the version of Burns Depression Checklist with 15 questions. BDC is one of the most commonly used resources when looking for depression symptoms. The Dr.

Burns depression checklist (BDC) is a trustworthy mood-measuring tool that accurately detects depression and ranks its severity. It has been used to determine the level of depression that each participant possess[11]. Participants had to rate the severity of several depressive symptoms they had experienced in the previous week, including the day of the survey, in order to be screened for depression using BDC. It is one of the most widely used rating scales for assessing depression. It is persistent and focuses on the specific symptoms of depression rather than the general symptoms of depression.

A person's overall BDC score is assessed by adding the intensity of each symptom that the person has given. In BDC, relation between overall score given by the participant and the level of depression they possess is given by the Table 1.

Total score	**Level of Depression**
0-4	**Minimal or no depression**
5-10	**Borderline depression**
11-20	**Mild depression**
21-30	**Moderate depression**
31-45	**Severe depression**

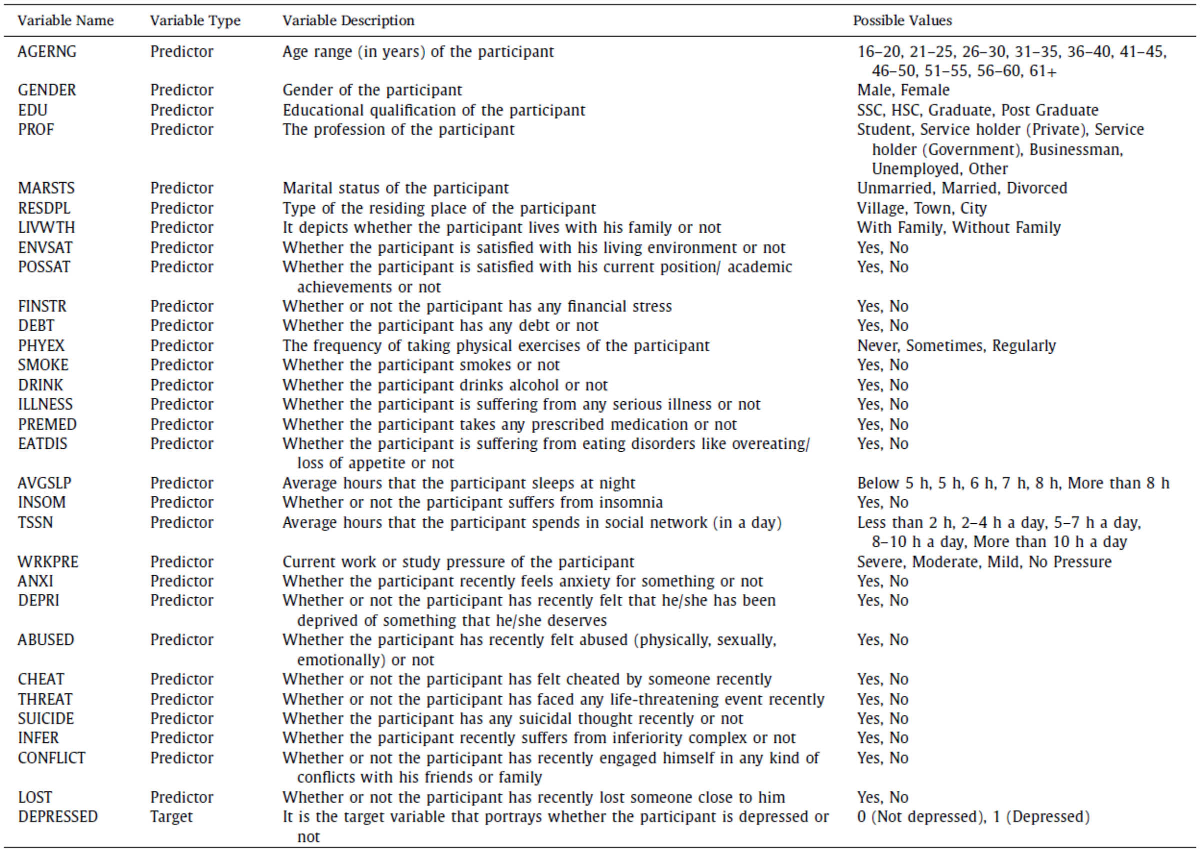
Table 1: Accuracy obtained for different models using t-SNE

3.2 Data Description

The survey's dataset contains thirty predictor variables and one target variable. The Burns Depression Checklist (BDC) was used to create the target variable for each of the participants. The first 30 questions related to psychosocial, and socio-demographic information of the participants is given in fig 1.

3.3 Data Analysis

Among the 374 participants of the accumulated dataset,174 participants have been found depressed. As the dataset has been collected during the time of COVID-19 and the pandemic has created a psychosocial and socio-economic crisis all over the world, it may have triggered the increase of the prevalence of depression among the participants. Out of the



Variable Name	Variable Type	Variable Description	Possible Values
AGERNG	Predictor	Age range (in years) of the participant	16–20, 21–25, 26–30, 31–35, 36–40, 41–45, 46–50, 51–55, 56–60, 61+
GENDER	Predictor	Gender of the participant	Male, Female
EDU	Predictor	Educational qualification of the participant	SSC, HSC, Graduate, Post Graduate
PROF	Predictor	The profession of the participant	Student, Service holder (Private), Service holder (Government), Businessman, Unemployed, Other
MARSTS	Predictor	Marital status of the participant	Unmarried, Married, Divorced
RESDPL	Predictor	Type of the residing place of the participant	Village, Town, City
LIVWTH	Predictor	It depicts whether the participant lives with his family or not	With Family, Without Family
ENVSAT	Predictor	Whether the participant is satisfied with his living environment or not	Yes, No
POSSAT	Predictor	Whether the participant is satisfied with his current position/ academic achievements or not	Yes, No
FINSTR	Predictor	Whether or not the participant has any financial stress	Yes, No
DEBT	Predictor	Whether the participant has any debt or not	Yes, No
PHYEX	Predictor	The frequency of taking physical exercises of the participant	Never, Sometimes, Regularly
SMOKE	Predictor	Whether the participant smokes or not	Yes, No
DRINK	Predictor	Whether the participant drinks alcohol or not	Yes, No
ILLNESS	Predictor	Whether the participant is suffering from any serious illness or not	Yes, No
PREMED	Predictor	Whether the participant takes any prescribed medication or not	Yes, No
EATDIS	Predictor	Whether the participant is suffering from eating disorders like overeating/ loss of appetite or not	Yes, No
AVGSLP	Predictor	Average hours that the participant sleeps at night	Below 5 h, 5 h, 6 h, 7 h, 8 h, More than 8 h
INSOM	Predictor	Whether or not the participant suffers from insomnia	Yes, No
TSSN	Predictor	Average hours that the participant spends in social network (in a day)	Less than 2 h, 2–4 h a day, 5–7 h a day, 8–10 h a day, More than 10 h a day
WRKPRE	Predictor	Current work or study pressure of the participant	Severe, Moderate, Mild, No Pressure
ANXI	Predictor	Whether the participant recently feels anxiety for something or not	Yes, No
DEPRI	Predictor	Whether or not the participant has recently felt that he/she has been deprived of something that he/she deserves	Yes, No
ABUSED	Predictor	Whether the participant has recently felt abused (physically, sexually, emotionally) or not	Yes, No
CHEAT	Predictor	Whether or not the participant has felt cheated by someone recently	Yes, No
THREAT	Predictor	Whether or not the participant has faced any life-threatening event recently	Yes, No
SUICIDE	Predictor	Whether the participant has any suicidal thought recently or not	Yes, No
INFER	Predictor	Whether the participant recently suffers from inferiority complex or not	Yes, No
CONFLICT	Predictor	Whether or not the participant has recently engaged himself in any kind of conflicts with his friends or family	Yes, No
LOST	Predictor	Whether or not the participant has recently lost someone close to him	Yes, No
DEPRESSED	Target	It is the target variable that portrays whether the participant is depressed or not	0 (Not depressed), 1 (Depressed)

Figure 1: Variables for predicting depression

total population who have participated in the survey 37 percent resides in village and the remaining resides in city and town. About half of the total population who has participated in the survey were graduates. Also more than half that is 58.8 percent were students. About 77.4 percent of the total participants were unmarried.

4 Feature selection techniques

Only a few variables in the dataset are important for generating the machine learning model, and the remaining features are either redundant or unnecessary. If we populate the dataset with all of these redundant and irrelevant information, the model's overall performance and accuracy will suffer. As a result, identifying and selecting the most appropriate characteristics from the data, as well as removing unnecessary or less important information, is critical, which is accomplished through feature selection in machine learning.

4.1 SelectKBest

The SelectKBest() function in Sklearn allows you to use a univariate statistical test to select a set of features. The statistical test function looks for features that are most closely related to the target feature. The K-best features were chosen using a chi-square test-based technique in this research. In statistics, the Chi-squared test is used to determine if two occurrences are independent. We utilise it in feature selection to see if the occurrence of a particular characteristic and the target are independent or not.

4.2 Minimum redundancy and maximum relevance (mRMR)

This method selects a subset of features that have the highest correlation with the class (output) and the lowest correlation with each other. It uses the minimal-redundancy-maximal-relevance criterion, which is based on mutual information, to rank characteristics. The Pearson correlation coefficient can be used to calculate correlation between features and the F-statistic can be used to calculate correlation with the class (relevance) (redundancy). Then, using a greedy search to maximise the objective function, which is a function of relevance and redundancy, features are chosen one by one. The MID and MIQ objective functions, which reflect the difference or quotient of relevance and redundancy, respectively, are two extensively utilised types of objective function.

4.3 Boruta Algorithm

The Boruta algorithm wraps the Random Forest classification algorithm. This approach claims to pick critical features by eliminating irrelevant attributes iteratively. The following are the steps of the Boruta algorithm:

- Step I: Duplicates of all features to expand the dataset.
- Step II: The values of the replicated features are shuffled to produce shadow features. Shuffling is used to get rid of their connections to the target variable.

- Step III: The expanded dataset is subjected to a random forest technique, and Z-scores are calculated.
- Step IV: The MZSA (Maximum Z-score among Shadow Attributes) is then determined.

4.4 Synthetic minority oversampling technique (SMOTE)

The dataset must be balanced to improve the minority class's forecast accuracy. To address the problem of class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) is applied. SMOTE uses feature space to generate synthetic samples of the minority class. The synthetic samples are introduced along the line that runs parallel to each minority class sample and its K-nearest minority class sample neighbours. First, the difference between the feature vector of the minority class instance under examination and its nearest neighbour is multiplied by a random number between 0 and 1 to create a synthetic instance. The multiplied result is then added to the feature vector in question, resulting in a synthetic instance of minority class. Using an unbalanced dataset to train a classifier results in biased and erroneous predictions. The percentages of depressed and non-depressed subjects in the training datasets are 66.87 percent and 33.13 percent, respectively. SMOTE was utilised to solve the problem of class imbalance in the training datasets because they were severely imbalanced.Table

5 Machine learning techniques for depression detection

In order to predict the existence of depression, this study has used six different machine learning classifiers, namely: K-Nearest Neighbor (KNN), Adaptive Boosting (AdaBoost), Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), Bagging classifier.

5.1 K-nearest neighbor (KNN)

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. KNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using KNN algorithm. KNN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. KNN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

5.2 Adaptive boosting (AdaBoost)

Boosting is an ensemble modelling strategy that aims to create a strong classifier out of a large number of weak ones. It is accomplished by constructing a model from a sequence of weak models. To begin, a model is created using the training data. The second model is then created, which attempts to correct the faults in the previous model. This approach is repeated until either the entire training data set is properly predicted or the maximum number of models has been added.

AdaBoost is a type of ensemble learning (sometimes known as "meta-learning") that was designed to improve the efficiency of binary classifiers. AdaBoost employs an iterative strategy to improve poor classifiers by learning from their mistakes.

5.3 Gradient Boost(GB)

Gradient Boosting (GB) classifiers combine a group of weak models to produce new models in a sequential manner. Each new model aims to reduce the loss function to the smallest possible value. The loss function is computed by GB using the gradient descent method. To avoid difficulties with overfitting, boosting should be halted as soon as possible using stopping criteria. As a stopping criteria, a maximum number of models constructed or a threshold on predicted accuracy can be employed .

5.4 Extreme gradient boosting (XGBoost)

XGBoost stands for “Extreme Gradient Boosting”. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements Machine Learning algorithms under the Gradient Boosting framework. It provides a parallel tree boosting to solve many data science problems in a fast and accurate way. It became popular in the recent days and is dominating applied machine learning and Kaggle competitions for structured data because of its scalability. XGBoost is an extension to gradient boosted decision trees (GBM) and specially designed to improve speed and performance.

5.5 Bagging Classifier

Bagging was created using the ideas of bootstrapping and aggregation. Bootstrap datasets are created from the training dataset in the Bagging classifier. Following that, each of these bootstrap datasets is utilised to train several classifiers. Finally, the outcomes of these classifiers are combined to get the final forecast. Misleading training items are frequently avoided in the bootstrap dataset. In many cases, combining the classifiers yields better results than using a single classifier. Because both of these characteristics are integrated in the Bagging classifier, it often outperforms other classifiers.

6 Results and discussion

By applying different feature selection techniques, accuracies of all of these classifiers have been increased dramatically. While using the MRMR feature selection technique, the GradientBoost has outperformed the other classifiers in terms of accuracy. It has achieved an accuracy of 80 percent. By applying the SelectKBest feature selection technique, the accuracies of the other classifiers namely, KNN,AdaBoost, GB, XGBoost and Bagging Classifier are 65.33 percent, 72 percent, 69.3 percent, 72 percent, and 73.33 percent respectively.

In the case of using the mRMR feature selection technique, KNN, AdaBoost, GB, XGBoost and Bagging classifiers have attained accuracies of 74.6 percent, 72 percent, 80 percent, 73.3 percent and 76 percent, respectively.

Using the Boruta feature selection technique, AdaBoost has shown superior performance

Model name	**Accuracy (in percentage)**
K-nearest neighbor (KNN)	**65.33**
AdaBoost	**72**
Gradient Boost	**69.3**
Extreme gradient boosting	**72**
Bagging	**73.33**

Table 2: Accuracy obtained for different models using SelectKBest

Model name	**Accuracy (in percentage)**
K-nearest neighbor (KNN)	**74.6**
AdaBoost	**72**
Gradient Boost	**80**
Extreme gradient boosting	**73.3**
Bagging	**76**

Table 3: Accuracy obtained for different models using mRMR

than the other classifiers in terms of accuracy. Here, the achieved accuracies of KNN, AdaBoost, GB, XGBoost and Bagging classifiers are 72 percent, 74.6 percent, 73.33 percent, 74 percent and 76 percent respectively.

Model name	**Accuracy (in percentage)**
K-nearest neighbor (KNN)	**72**
AdaBoost	**74.6**
Gradient Boost	**73.33**
Extreme gradient boosting	**74**
Bagging	**76**

Table 4: Accuracy obtained for different models using Boruta

These models' Area Under Curve (AUC) values have also been determined. If the AUC value of a model is 1, it is assumed that it is a perfect model or classifier. When a model's AUC value is 0.5, it can't tell the difference between samples from different classes. As a result, a model with a higher AUC value is always preferred. The AUC, Precision, and F1-score of the classifiers have all improved as a result of the given feature selection strategies. The Receiver Operator Characteristic (ROC) curve is used to highlight the trade-off between model sensitivity and specificity. It's a two-dimensional graph with the False Positive Rate on the x-axis and the True Positive Rate on the y-axis. The closer a classifier's ROC curve is to the top-left corner of the graph, the higher the classifier's performance. The ROC curves of these classifiers utilising different feature selection strategies are shown in Fig. 2 (a)–(d). After employing feature selection approaches, the ROC curves of the classifiers have shifted closer to the graph's upper left corner, as seen in Fig. 2.

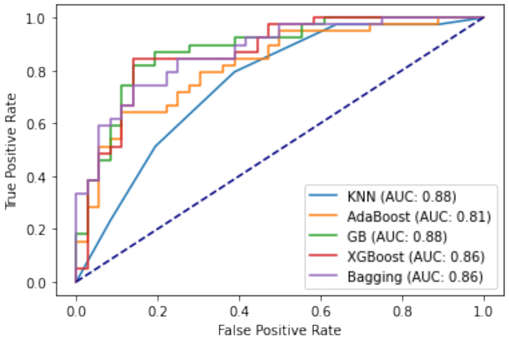


Figure 2: ROC Curve without using feature selection

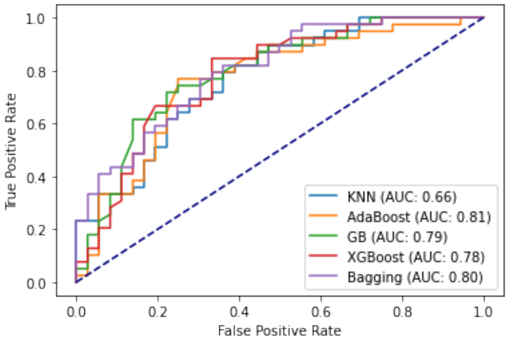


Figure 3: ROC Curve using SelectKBest

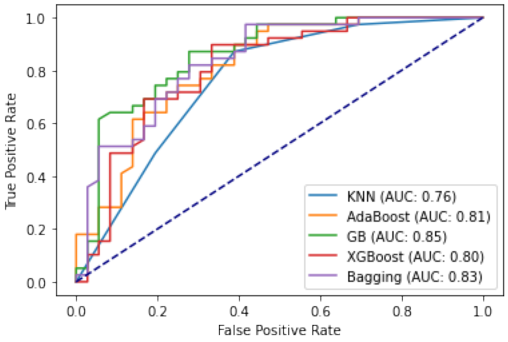


Figure 4: ROC Curve using mRMR

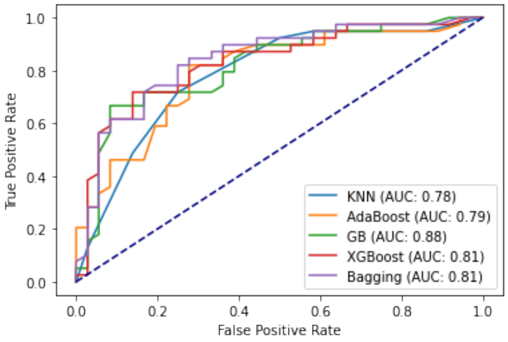


Figure 5: ROC Curve using Boruta

7 Conclusion

Several factors can contribute to a person's depression. The goal of this study was to identify the most frequent causes of depression. To begin with, a dataset comprising thirty socio-demographic and psychological characteristics from 374 people was developed to screen for depression. The most relevant demographic and psychosocial characteristics that lead to depression have been retrieved using various feature selection strategies. These feature selection strategies have improved the classifiers' training speed as well as their ability to detect sadness more precisely. This study used six different machine learning classifiers to determine the presence of depression. By observing the outcomes of various models presented in this study, it can be confirmed that the GradientBoost classifier with the MRMR feature selection technique is almost the perfect model to predict depression among the participants. It has obtained an accuracy of 80

References

[1] Tadesse MM, Lin H, Xu B, Yang L. Detection of depression-related posts in reddit social media forum. IEEE Access. 2019 Apr 4;7:44883-93.

[2] Priya, A. , Garg, S. , Tigga, N.P. , 2020. Predicting anxiety, depression and stress in modern life using machine learning algorithms. Procedia Comput. Sci. 167, 1258–1267 .

[3] Sau, A. , Bhakta, I. , 2019. Screening of anxiety and depression among seafarers using ma- chine learning technology. Inf. Med. Unlocked 16, 100228

[4] Hatton, C.M. , Paton, L.W. , McMillan, D. , Cussens, J. , Gilbody, S. , Tiffin, P.A. , 2019. Pre- dicting persistent depressive symptoms in older adults: a machine learning approach to personalised mental healthcare. J. Affect. Disord. 246, 857–860 .

[5] Vieira S, Pinaya WH, Mechelli A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. Neuroscience Biobehavioral Reviews. 2017 Mar 1;74:58-75.

[6] Orru G, Pettersson-Yeo W, Marquand AF, Sartori G, Mechelli A. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neuroscience Biobehavioral Reviews. 2012 Apr 1;36(4):1140-52.

[7] Orrù G, Monaro M, Conversano C, Gemignani A, Sartori G. Machine learning in psychometrics and psychological research. Frontiers in psychology. 2020 Jan 10;10:2970.

[8] https://www.ndtv.com/kerala-news/mental-depression-high-among-people-of-kerala-survey-1627899

[9] Sagna A, Gallo JJ, Pontone GM. Systematic review of factors associated with depression and anxiety disorders among older adults with Parkinson's disease. Parkinsonism related disorders. 2014 Jul 1;20(7):708-15.

[10] Cole MG, Dendukuri N. Risk factors for depression among elderly community subjects: a systematic review and meta-analysis. American journal of psychiatry. 2003 Jun 1;160(6):1147-56.

[11] Vink D, Aartsen MJ, Schoevers RA. Risk factors for anxiety and depression in the elderly: a review. Journal of affective disorders. 2008 Feb 1;106(1-2):29-44.