DEPRESSION DETECTION USING SOCIO-DEMOGRAPHIC FEATURES AND PSYCHOSOCIAL FACTORS

Report of Internship

TATA CONSULTANCY SERVICES (TCS)

15-Nov-2021 to 27-May-2022

Submitted by

Mrs. THASNEEM VAZIM

Intern Emp ID: 2170620

Final-year Postgraduate student

Centre for Artificial Intelligence

THANGAL KUNJU MUSALIAR COLLEGE OF ENGINEERING KERALA

Under the guidance of

RAJEEV AZHUVATH (TCS Mentor ID: 120914)

&

Prof. SUMOD SUNDAR (TKMCE)

MAY 2022

ACKNOWLEDGEMENT

A successful project is a fruitful culmination of efforts by many people, some directly involved and some others indirectly, by providing support and encouragement. Firstly I would like to thank the almighty for giving me the wisdom and grace to make my project a successful one. I thank him for steering me to the shore of fulfilment under his protective wings.

With a profound sense of gratitude, I would like to express my heartfelt thanks to my mentor **Rajeev Azhuvath, Tata Consultancy Services (TCS)**, for his expert guidance, constant support and cooperation.

I express my sincere gratitude to **Dr. T A Shahul Hameed**, Principal of TKMCE and **Dr. Imthias Ahamed**, Professor and Head of the Department, Centre for Artificial Intelligence, TKMCE, for their immense encouragement. I would like to thank my college guide **Prof. Sumod Sundar**, Assistant Professor, Centre for Artificial Intelligence, TKMCE, for his expert guidance and cooperation.

I also express my thanks to my loving parents, brother and friends, for their support and encouragement in the successful completion of this project work.

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 use 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 percent.

Contents

List of Figures

Chapter 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].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.

Kerala is situated on the southwestern coast of India, a country in South Asia. According to a recent survey, almost 9 percent of individuals in Kerala suffer from mental depression, which affects people of all ages. It was also discovered that one out of every eight people, or 12.43 percent of those surveyed, required psychiatric help. According to the poll, depression was the most common mental disease among those questioned. Around 9 percent of people had depression, 0.29 percent had schizophrenia, and 0.27 percent had bipolar illness, according to the study. According to the survey, 75 percent of respondents are receiving treatment for mental health disorders, while 25 percent are still waiting for therapy. So, it is necessary to screen and identify depressed persons at an early stage[7].

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.

Chapter 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]

Chapter 3

Methodology

The flowchart of the proposed model is shown in fig 3.1.

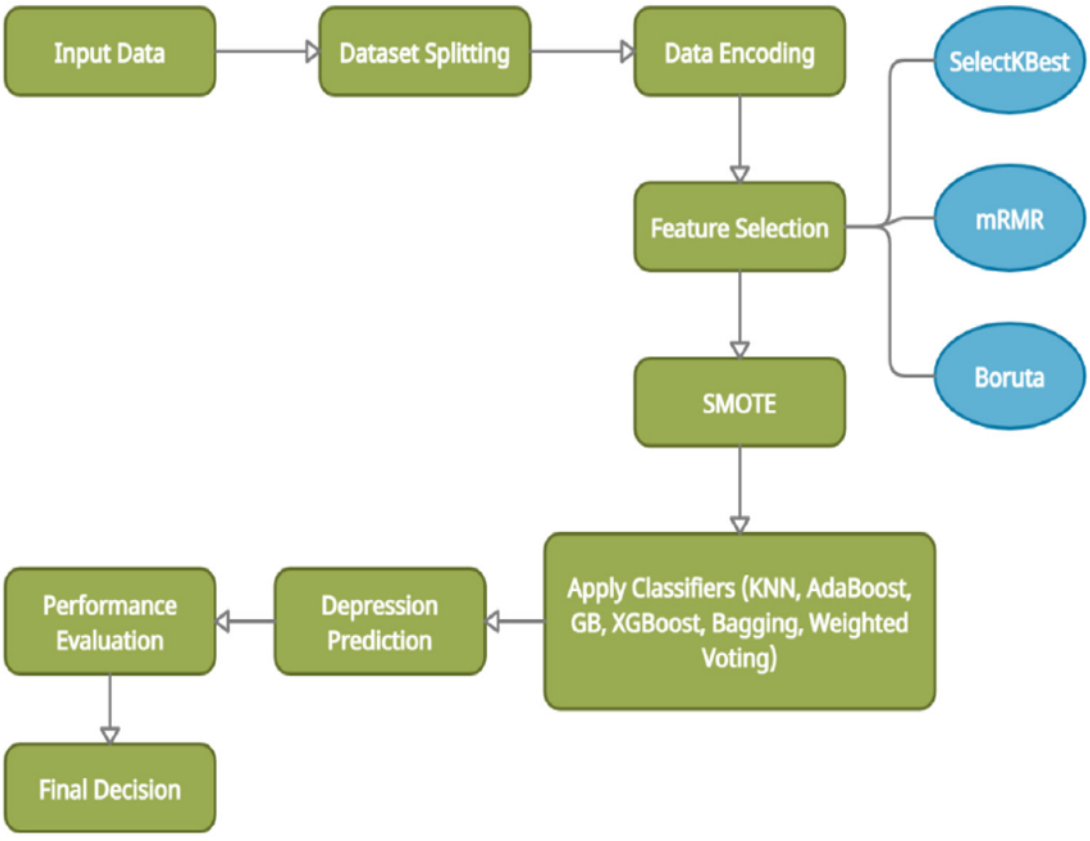

Figure 3.1: Flowchart of proposed model

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.

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 3.2.

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 total population who have participated in the survey 37 percentage of the people 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.

3.4 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

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 3.2: Variables for predicting depression

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.

3.4.1 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.

3.4.2 Minimum redundancy and maximum relevance (mRMR)

This algorithm tends to select a subset of features having the most correlation with the class (output) and the least correlation between themselves . It ranks features according to the minimal-redundancy-maximal-relevance criterion which is based on mutual information. The F-statistic can be used to calculate correlation with the class (relevance) and the Pearson correlation coefficient can be used to calculate correlation between features (redundancy). Thereafter, features are selected one by one by applying a greedy search to maximize the objective function, which is a function of relevance and redundancy. Two commonly used types of the objective function are MID and MIQ representing the difference or the quotient of relevance and redundancy, respectively.

3.4.3 Boruta Algorithm

The Boruta algorithm works as a wrapper around the Random For- est classification algorithm . This method aims to select important features by iteratively removing the irrelevant attributes. The steps of Boruta algorithm are given below:

Step I: The dataset is extended by creating duplicates of all the features.

Step II: Shadow features are created by shuffling the values of the duplicated features. Shuffling is performed for removing their corre- lations with the target variable.

Step III: A random forest algorithm is applied to the extended dataset, and Z -scores are computed.

Step IV: The Maximum Z -score among the Shadow Attributes (MZSA) is then detected.

Step V: If a feature's importance is remarkably less than MZSA, then it is permanently removed from the dataset. On the other hand, if a feature's importance is remarkably greater than MZSA, then it is kept in the dataset.

Step VI: Discard the shadow features from the dataset.

Step VII: Repeat the steps until there is no unimportant feature in the dataset or for a predefined number of iterations.

3.4.4 SMOTE

Training a classifier with an imbalanced dataset leads to biased and inaccurate predictions. In the training datasets, the percentages of depressed and not depressed participants are 66.87 percent and 33.13 percent respectively. As the training datasets are highly imbalanced, SMOTE has been used to remove their class imbalance problem.

3.5 Training and testing for predicting depression

In this step, the classifiers namely, KNN, AdaBoost, GB, XGBoost and Bagging are trained with the training datasets. Following the training of these classifiers, each of them has been used to predict the depression of the participants of the test datasets.

3.6 Implementation

3.6.1 Dataset splitting

Firstly, the obtained dataset has been split into training and test data. This study has used 80 percent data of the dataset as training data. And the rest 20 percent data of the dataset has been used for testing purposes.

3.6.2 Data encoding

After the completion of the Dataset Splitting technique, Data En- coding is performed on the obtained training and test datasets. Using numeric data, the majority of machine learning algorithms demonstrate better results. In the Data Encoding step, the categorical data of the training and test datasets have been converted into their numeric counterpart using the Label Encoder of the Scikit-learn library.

3.6.3 Modifying the training and testing dataset using feature selection

The presence of irrelevant features degrades the classifiers' effi- ciency. For extracting the relevant and necessary features from the dataset, three feature selection techniques have been used in this study separately. Both the SelectKBest and mRMR feature selection techniques have chosen fifteen predictor variables separately for performing classification efficiently. And using the Boruta feature selection algorithm, thirteen most relevant predictor variables have been extracted from the total 6 predictor variables of the dataset. Figure 3.3 shows the list of the selected features using these three feature selection techniques.

Feature Selection Technique	Total Features	Selected Features
SelectKBest	15	DEPRI, INFER, POSSAT, ANXI, ABUSED, CHEAT, CONFLICT, FINSTR, SUICIDE, ENVSAT, INSOM, THREAT, LOST, DEBT, EATDIS
mRMR	15	DEPRI, POSSAT, ANXI, INFER, ENVSAT, CHEAT, FINSTR, ABUSED, CONFLICT, SUICIDE, LOST, INSOM, THREAT, WRKPRE, DEBT
Boruta	13	ENVSAT, POSSAT, FINSTR, INSOM, ANXI, DEPRI, ABUSED, CHEAT, THREAT, SUICIDE, INFER, CONFLICT, LOST

Figure 3.3: Selected features using feature selection techniques.

Chapter 4

Results and discussion

By applying different feature selection techniques, accuracies of all of these classifiers have been increased dramatically. While using the SelectKBest feature selection technique, the AdaBoost has outperformed the other classifiers in terms of accuracy. It has achieved an accuracy of 92.56 percent. By applying the SelectKBest feature selection technique, the accuracies of the other classifiers namely, KNN, GB, XGBoost and Bagging 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 than the other classifiers in terms of accuracy. Here, the achieved accuracies of KNN, AdaBoost, GB, XGBoost and Bagging are 72 percent, 74.6 percent, 73.33 percent, 74 percent and 76 percent respectively.

Fig. 2 (a)–(d) shows the ROC curves of these classifiers using different feature selection techniques. Fig. 2 reveals that the ROC curves of the classifiers have moved closer to the graph's upper left corner after applying feature selection techniques.

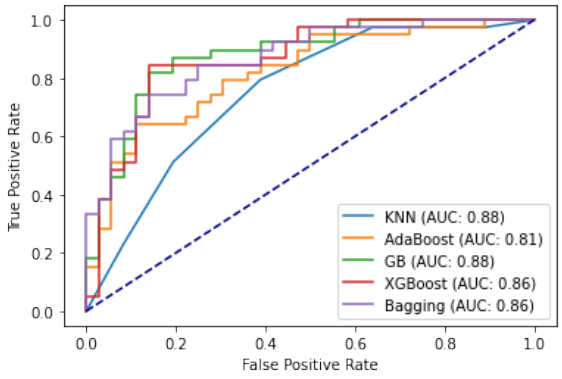

Figure 4.1: ROC Curve without using feature selection

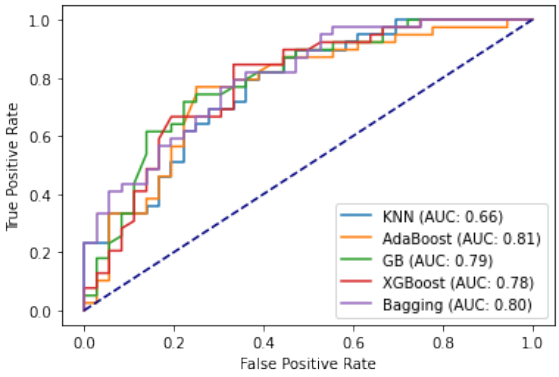

Figure 4.2: ROC Curve using SelectKBest

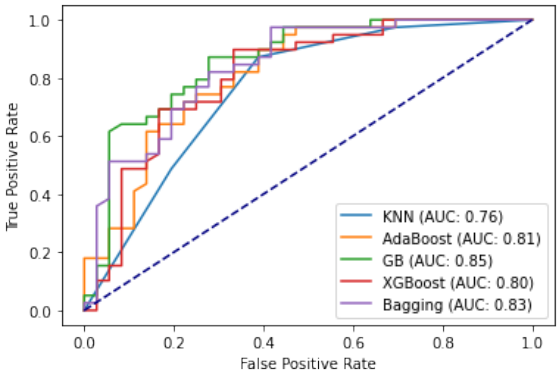

Figure 4.3: ROC Curve using MRMR

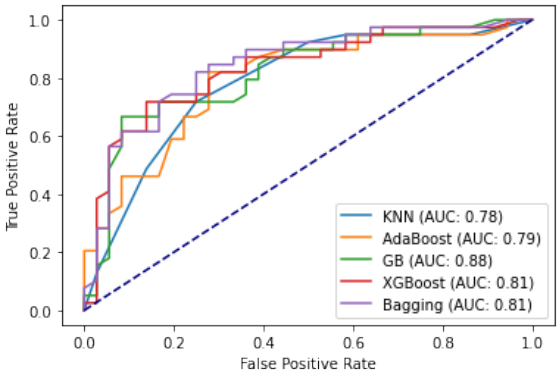

Figure 4.4: ROC Curve using Boruta

Chapter 5

Conclusion

Various factors can play roles in forming depression in a person. This study has tried to find out the most common factors that cause depression. Firstly, a dataset has been created, consisting of thirty socio- demographic, and psychosocial factors of 604 participants to screen depression. Different feature selection techniques have extracted the most important demographic, and psychosocial factors responsible for forming depression. These feature selection techniques have not only boosted the training speed of the classifiers but also helped the classifiers to screen depression more precisely. To ascertain the presence of depression, this research has used six different machine learning classifiers. 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 percent.

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.