AUTOMATED DEPRESSION DETECTION USING TABNET CLASSIFIER

A Project Report

Submitted by

Ms. THASNEEM VAZIM

REG NO: TKM20MEAI15

SEMESTER: IV

In partial fulfillment for the award of the degree of

MASTER OF TECHNOLOGY
IN
Mechanical Engineering (Artificial Intelligence)

Under the guidance of
Prof. SUMOD SUNDAR

Thangal Kunju Musaliar College of Engineering
Kerala

JULY 2022

DECLARATION

I undersigned hereby declare that the project report "Automated depression detection using Tabnet classifier", submitted for partial fulfillment of the requirements for the award of degree of Master of Technology of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me under supervision of Prof. Sumod Sundar. This submission represents my ideas in my own words and where ideas or words of others have been included, I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

Place: Kollam
Date: THASNEEM VAZIM

Thangal Kunju Musaliar College of Engineering
Centre for Artificial Intelligence

C E R T I F I C A T E

This is to certify that, this report titled ***AUTOMATED DEPRESSION DETECTION USING TABNET CLASSIFIER*** is a bonafide record of the **Project** presented by **THASNEEM VAZIM (TKM20MEAI15)**, under our guidance and supervision, in partial fulfillment of the requirements for the award of the degree, **M.Tech in Mechanical Engineering (Artificial Intelligence)** in **APJ Abdul Kalam Technological University** .

Project coordinator & Internal supervisor

Prof. Sumod Sundar
Assistant Professor
Centre for Artificial Intelligence

Head of the Department

Dr. Imthias Ahamed
Professor & HOD
Centre for Artificial Intelligence

Internal examiner

External examiner

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 for making my project a successful one. I thank him for steering me to the shore of fulfillment under his protective wings.

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 constant support and encouragement throughout the project work.

I would like to express my heartfelt thanks to our project coordinator cum internal supervisor **Prof. Sumod Sundar**, Assistant Professor, Centre for Artificial Intelligence, TKMCE, for his expert guidance and cooperation. With a profound sense of gratitude, I would like to thank **Dr. Santhi Natarajan**, Honorary Professor, for her immense encouragement. I would like to express my gratitude to **Mr. Rajeev Azhuvath**, Mentor, Tata Consultancy Services (TCS), for his expert guidance, and cooperation. I also extend my thanks to the entire faculty and staff members of the Centre for AI, TKMCE, who has encouraged me throughout this work.

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

THASNEEM VAZIM

Abstract

Depression, also 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% of the global population. Multiple studies have shown that early detection and treatment of depression can reduce the disorder's harmful effects. This study looked at TabNet, five machine learning classifiers, namely K-nearest neighbor (KNN), Adaptive boosting (AdaBoost), Gradient Boosting(GB), Extreme gradient boosting (XGBoost), and Bagging classifiers. These models used various socio- demographic and psychological data to determine whether or not a person is depressed. Natural Language Processing (NLP) techniques and machine learning methodologies have been used to train the data and evaluate the efficiency of the proposed model. The Synthetic Minority Oversampling Technique (SMOTE), which minimizes the class imbalance of the training data was utilized to improve accuracy in predicting depression. The feature extraction techniques, namely SelectKBest, Minimum redundancy and maximum relevance(mRMR), and Boruta algorithm, were used for extracting the most essential socio-demographic and psychosocial factors responsible for forming a depression. The TabNet was found to be the perfect model to predict depression among the participants, with an accuracy of 91.78%.

Contents

List of Figures

List of Tables

ABBREVIATIONS

AUC	Area Under the ROC Curve
BDC	Burns Depression Checklist
DNN	Dense Neural Network
ERDE	Early Risk Detection Error.
GB	Gradient Boosting
KNN	K-Nearest Neighbor
MID	Mutual Information Difference
MIQ	Mutual Information Quotient
MLP	Multi Layer Perceptron
mRMR	Minimum Redundancy and Maximum Relevance
NLP	Natural Language Processing
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Oversampling TEchnique
LSTM	Long Short Term Memory
RNN	Recurrent Neural Network
WHO	World Health Organisation
XGBoost	eXtreme Gradient Boosting

Chapter 1

INTRODUCTION

Depression is a frequent mental health condition that negatively affects a person's moods, thought processes, and behaviours, as well as their capacity to operate properly. Because of the stigma associated with mental health, most persons with depression prefer to hide their symptoms and avoid seeking professional care.In many countries, depression is still undiagnosed and untreated, which can lead to negative self-perception and, in the worst- case scenario, suicide. Depression has been considered as a significant factor that leads to suicidal thoughts. Despite this, many people who suffer from depression are not treated for various reasons. Multiple studies have shown that early detection and treatment of depression can reduce the disorder's harmful effects. Timely identification of depressed symptoms, followed by assessment and therapy, can significantly improve the chances of controlling symptoms and attenuate harmful effects on personal, economic, and social life. Depression was named the most preventable condition by the Institute of Medicine Committee on the Prevention of Mental Disorders.

According to the study conducted by the World Health Organization (WHO), 300 million people worldwide suffer from depression. In 2015, it was projected that four percent of the world's population was depressed . Suicide is the second leading cause of death among those aged 15 to 29. Suicide is most commonly caused by depression. Every year, over 80,000 people die by suicide. Depression analysis is the method of determining whether or not someone is depressed based on their social media textual activity. The problem of detecting depression from social media has been posed as a classification problem in the Natural Language Processing domain (NLP). In this paper, we look at NLP techniques that can successfully extract information from textual material in order to improve depression detection. To create document representations, these NLP techniques extract different features.

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. Determining depression symptoms, on the other hand, is difficult and time-consuming. Current methods rely mostly on clinical interviews and questionnaire surveys conducted by hospitals or organisations , with psychological evaluation tables used to create mental disorder predictions. This method relies primarily on one-on-one surveys to identify depression.

In psychological analysis and psychometrics, statistical inferences have been employed for decades. Machine Learning (ML) in psychometrics attracted media attention after the Cambridge Analytica affair. Machine learning is a collection of algorithms that can automat-

ically recognize patterns in data and utilize those patterns to forecast future data.Machine learning (ML) has been introduced into the medical field to provide diagnostic tools capable of enhancing accuracy and precision while minimizing laborious tasks that require human intervention. Psychologists extensively use least-squares regression techniques to find patterns in data. In contrast to statistical inference, ML analysis of experimental data is model agnostic and primarily focuses on prediction rather than inference.Due to replicability issues with statistical inference, researchers in psychometrics and psychological analysis are turning to machine learning from statistical inferences. However, when compared to other fields, ML is not often employed in the study of psychological trials (e.g., genetics). Moreover, machine learning can help with time-consuming and complex jobs in the healthcare field. These innovative technologies help save money and improve therapeutic outcomes by accelerating good 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. These variables can be used to construct an automated depression prediction system. With the rapid growth of information and technology, using 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 presence of depression in an individual, the significant elements that cause depression, and the best classifier for detecting depressed people.The present study uses a database to investigate the the use of TabNet which is a novel high-performance and interpretable canonical deep tabular data learning architecture, and five machine learning models along with feature extraction techniques to predict the presence of depression in a participant using socio-demographic and psycho-social features.

Chapter 2

RELATED WORKS

Tadesse MM et al.investigated the performance of the single feature and combination feature sets to measure depression symptoms using various text classification algorithms[1].According to the findings, good feature selection and diverse feature combinations improved prediction performance. In the study, the MLP classifier achieved 91% accuracy and the 0.93 F1 scores earned the most significant performance for identifying depression on Reddit social media, demonstrating the strength and utility of integrated characteristics.

Priya et al. used machine learning algorithms to collect data on anxiety, stress, and depression to diagnose these mental illnesses[2].They discovered that the Naive Bayes classifier predicted depression the best, with an accuracy of 85.50%.

In their research, Sau et al.focused on the senior population[3].They used ten 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.

Hatton et al.used 284 senior individuals' psychometric and demographic data to predict the prevalence of depression[4].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.

Viera S et al. used the online DASS42 tool to predict five severity levels of anxiety, depression, and stress using eight different machine learning models[5].These methods are classified into four types: Bayes, neural networks, lazy methods, and tree methods.The final methodology combines the K-star and random forest techniques.The hybrid strategy improved single algorithm accuracy and took 30 to 45 minutes to complete, whereas single algorithms only took five minutes.

Marcel Trotzek et al use machine learning models based on messages on a social platform to detect depression early[6].A convolutional neural network based on different word embeddings is specifically evaluated and compared to a classification based on user-level linguistic metadata. In a current early detection task, an ensemble of both approaches achieves state-of-the-art results. Furthermore, the currently popular ERDE score as a metric for early detection systems is thoroughly examined, and its shortcomings in the context of shared tasks are demonstrated.The original score is compared to a slightly modified metric.Finally, a new word embedding was trained and evaluated on a large corpus of the same domain as the described task.

Ezekiel Victor et al developed AiME, a framework capable of detecting depression with minimal human intervention (Artificial Intelligence Mental Evaluation)[7]. AiME consists of a brief human-computer interactive evaluation and artificial intelligence, specifically deep learning, and can predict whether a participant is depressed or not with satisfactory performance. Because of its ease of use, this technology can provide a viable tool for mental health professionals to identify symptoms of depression, allowing for faster preventative intervention.Furthermore, it may make it easier to interpret highly nuanced physiological and behavioral biomarkers of depression by providing a more objective assessment.

Shah FM et al proposed Bidirectional Long Short Term Memory (BiLSTM) with various word embedding techniques and metadata features, which yielded promising results[8]. Not only did they want to ensure that everything was in order, but they also wanted to ensure that we not only diagnose depressed consumers but also help them spend less time depressed. The disadvantage is that, even though the users are correctly classified, detection takes too long as depression. More work will be required to solve this problem.

Tadesse MM et al investigated the performance of the single feature and combination feature sets to measure depression indicators using various text classification algorithms[9].According to the findings, good feature selection and diverse feature combinations result in better prediction performance. In their study, the MLP classifier achieved 91 % accuracy and the 0.93 F1 scores achieved the highest performance degree for identifying the presence of depression on Reddit social media, demonstrating the power and utility of integrated characteristics.

Uddin MZ et al proposed a multimodal human depression prediction approach based on a one-hot approach on robust features based on describing depression symptoms and RNN[10].First, text data from young users was obtained from ung.no, a Norwegian public information channel aimed at young people. The one-hot method is then used after sequentially extracting words from different sentences and words representing depression symptoms. Furthermore, the one-hot features were used to train a deep RNN using the LSTM method to model two distinct emotional states: depression and non-depression. Finally, the trained RNN was used to predict the underlying emotional state in text data from unknown sensors.

Chapter 3

METHODOLOGY

The overview of the various experiments conducted in this study is shown in fig 3.1.

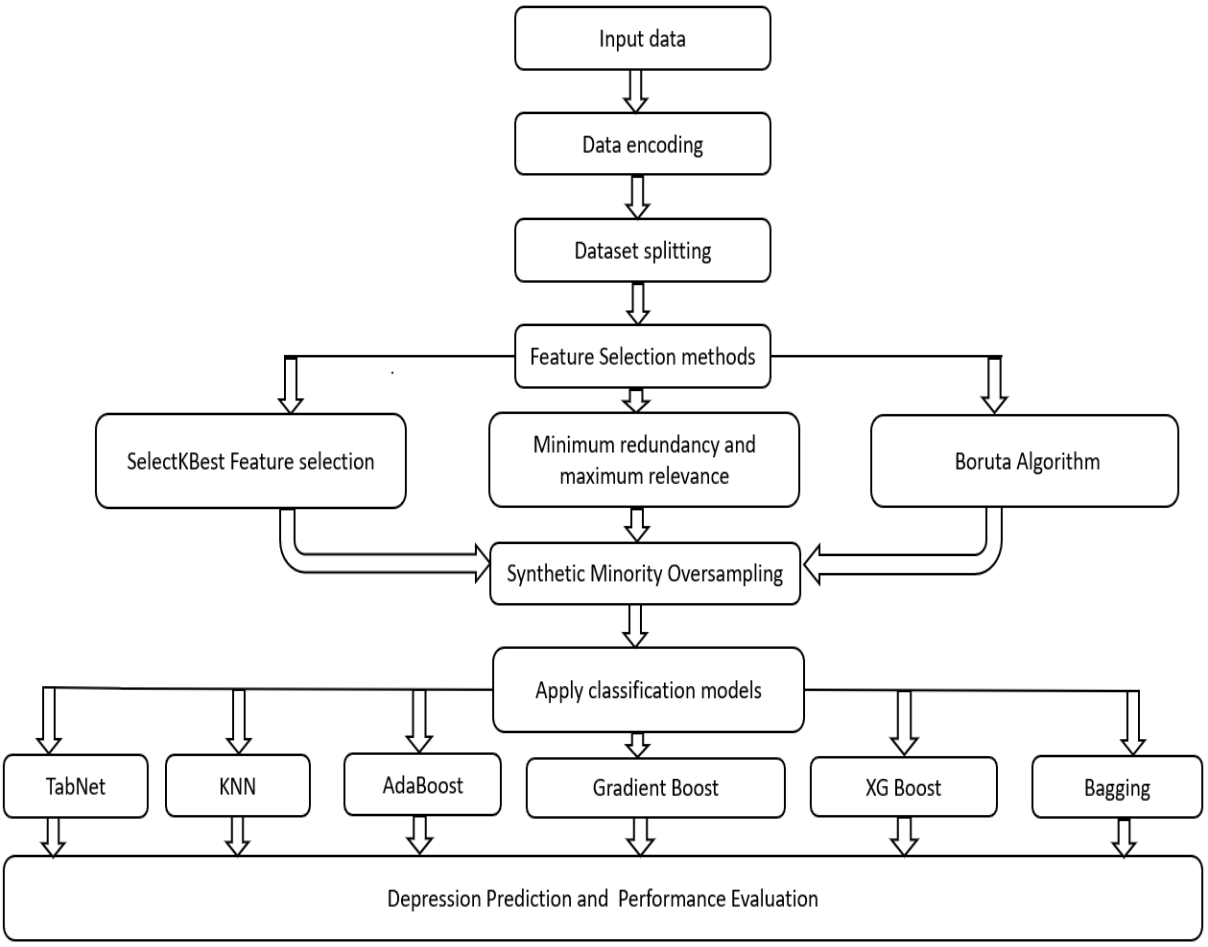

Figure 3.1: Flowchart of the experiments conducted during the study.

3.1 Data description

The work analyzes three datasets for predicting the depression among the participants.

The first dataset used was obtained from a survey conducted in the period between April 2022 and May 2022 to detect depression among Kerala citizens using socio-demographic features and psycho social factors.A questionnaire consisting of 45 questions was designed for the analysis. The first 30 questions were designed to manage the participants' psychosocial and socio-demographic data, and the last 15 questions were taken from the Burns Depression Checklist (BDC) and it consists of the responses of 374 participants.

The second dataset used for the work is taken from Github.The dataset is available in:https://github.com/ Sabab31/Depression Repository.git.This dataset was generated from conducting a survey among Bangladeshi citizens of different age ranges.A questionnaire with 55 questions was distributed among the participants.The first 30 questions were designed to elicit complex psychological and socio-demographic information from participants, while the latter 25 questions were drawn from the Burns Depression Checklist (BDC). The survey was carried out between April and August of 2020 and it consists of the responses of 604 individuals.

The third dataset is obtained by combining the above mentioned datasets,ie, Dataset 1 collected from Kerala citizens and Dataset 2 obtained from the github repository whose link is ;https://github.com/ Sabab31/Depression Repository.git followed by data augmentation to manage the overfitting of data.The dataset so obtained contained 1765 responses.

The three datasets contains thirty predictor variables and one target variable. The Burns Depression Checklist (BDC) was used to create the target variable for each participant.The Dr Burns depression checklist (BDC) is a reliable mood-measuring tool that accurately detects depression and ranks its severity. BDC is one of the most commonly used resources when looking for depression symptoms. 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. Participants had to rate the severity of several depressive symptoms they had experienced the previous week, including the day of the survey, to be screened for depression using BDC. A person's overall BDC score is assessed by adding the intensity of each symptom that the person has been given. Here a person is considered depressed if the overall score is greater than 10; otherwise, the participant is deemed to be a non-depressant. The first 30 questions related to psychosocial and socio-demographic information of the participants are given in Table 3.1.

3.2 Data analysis

During the analysis of the available dataset in Github it was found that 397 of the 604 individuals in the total sample were determined to be depressed. The prevalence of depression among survey participants is 66.87%, which is much higher than the actual prevalence of depression in Bangladesh's entire population.

Because the dataset was gathered during COVID-19, and the pandemic caused a psychological and socioeconomic catastrophe all across the world, it is possible that the subjects' prevalence of depression increased. The poll indicated that 34.27% of those who took part were not depressed.Female individuals are depressed more than male ones. The female participants have a depression prevalence of 69.80%, whereas the male participants have a

Feature description	**Probable values**
Age range (in years)	16–20, 21–25, 26–30, 31–35, 36–40, 40+
Gender	Male, Female
Educational qualification	SSC, HSC, Graduate, Post Graduate
Profession	Student, Unemployed, Other
Has financial stress or not?	Yes,No
Has financial stress or not?	Yes,No
Marital status	Unmarried, Married, Divorced
Do you live with your family or not	With Family,Without family
Satisfied with living environment or not	Yes,No
Satisfied with the current position or not	Yes,No
Has financial stress or not?	Yes,No
Has any debt or not?	Yes,No
The frequency of physical exercises	Never, Sometimes, Regularly
Do you smoke or not?	Yes,No
Do you drink alcohol or not?	Yes,No
Do you have any serious illness or not?	Yes,No
Do you take any prescribed medication	Yes,No
Suffering from any eating disorders?	Yes,No
Number of hours you sleeps at night	Below 5 h, 5 h, 6 h, 7 h, 8 h, More than 8 h
Do you suffer from insomnia?	Yes,No
Average hours spends in social media (in a day)	Less than 2 h, 2–4 h a day, 5–7 h a day,
Current work or study pressure	Severe, Moderate, Mild, No Pressure
Do you recently feels anxiety for something or not	Yes,No
Has recently felt deprived of something	Yes,No
Has recently felt abused or not	Yes,No
Has felt cheated by someone recently	Yes,No
Has faced any life-threatening	Yes,No
Has any suicidal thought recently or not	Yes,No
Suffers from inferiority complex	Yes,No
Recently engaged in any kind of conflicts.	Yes,No
Recently lost someone close to him/her	Yes,No
Target variable	0 (Not depressed), 1 (Depressed)

Table 3.1: Features used for predicting depression

depression prevalence of 64.40%. Participants with government occupations are the least impacted by depression, according to the profession. Only 40% of people in government jobs are depressed. Depression appears more frequently in business participants and the jobless. The frequency of depression among business participants is 90%. Seventy-five percent of jobless participants are depressed. The percentages of depression among students, private jobholders, and those with other jobs are 64.85%, 64.63%, and 70.59%, respectively.The divorced participants suffer the most from depression. The married participants had the lowest rate of depression. Depression affects 63.22 % of married, 65.95 percent of unmarried, and 100 % of divorced people, respectively. Participants who live in villages are less likely to suffer from depression. Depression is prevalent in 57.14 % of village people. Depression affects 68.75 % of people in town and 67.85% of participants in city, respectively.

During the analysis of dataset collected from Kerala citizens,174 participants have been found depressed from a total of 374 participants which constituted a value of 42.52%.Out of the total population who have participated in the survey 37% 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% were students.About 77.4% of the total participants were unmarried.

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.The percentage of male and female participants were 48% and 52% respectively.The percentage of depression among the females was much higher than males of about 54%,whereas the depression prevalence was 38% for males.51% of the total participants who resides in the village were found to be depressed.Depression is prevalent in 47% of the town population and 37% population in the city.The exploratory data analysis of the collected dataset is shown in fig 3.2 to 3.4.

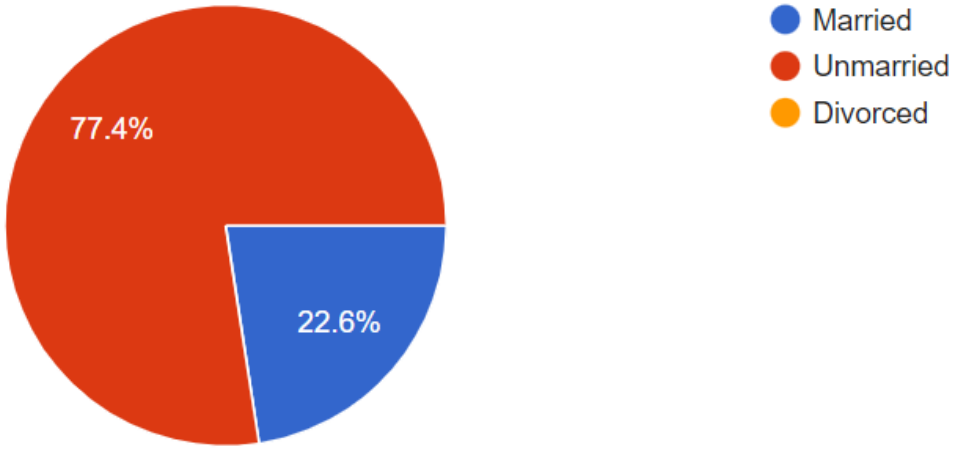

Figure 3.2: Marital status of participants from Kerala.

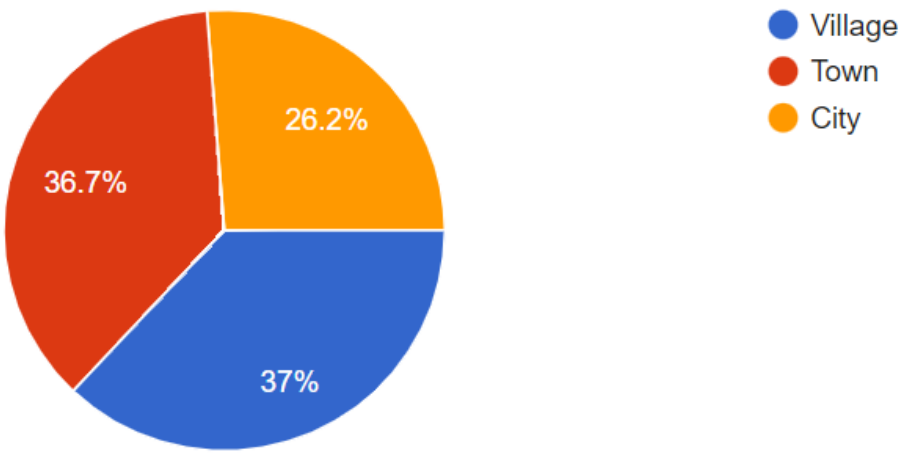

Figure 3.3: Residing place of the participant from Kerala.

3.3 Feature selection techniques

All the features in the dataset are not essential for generating the machine learning model, as some features may be either redundant or unnecessary[14].Using the feature selection technique, we select the data features that have the greatest impact on the target variable. The model's overall performance and accuracy will be affected if we populate the dataset with redundant and irrelevant information. As a result, identifying and selecting the most appropriate characteristics from the data and removing unnecessary or less important information is critical, which is accomplished through feature selection in machine learning.

3.3.1 SelectKBest

The SelectKBest() allows you to use a univariate statistical test to select a set of features[14]. The statistical test function looks for closely related parts to the target feature. The K-best features were chosen using a chi-square test-based technique in this research.The Chi-squared test is used in statistics to determine if two occurrences are independent. We utilized it in feature selection to see if the occurrence of a particular characteristic and the target are independent or not.

Eq. (1) can be used to calculate the chi-square score for n pairs of expected and observed frequencies.

$$X^2 = \Sigma \frac{(OFi - EFi)^2}{EFi} \tag{3.1}$$

Here, OFi is the observed frequency for the i-th value of the feature F, and EFi is the expected frequency for the i-th value of the feature F.

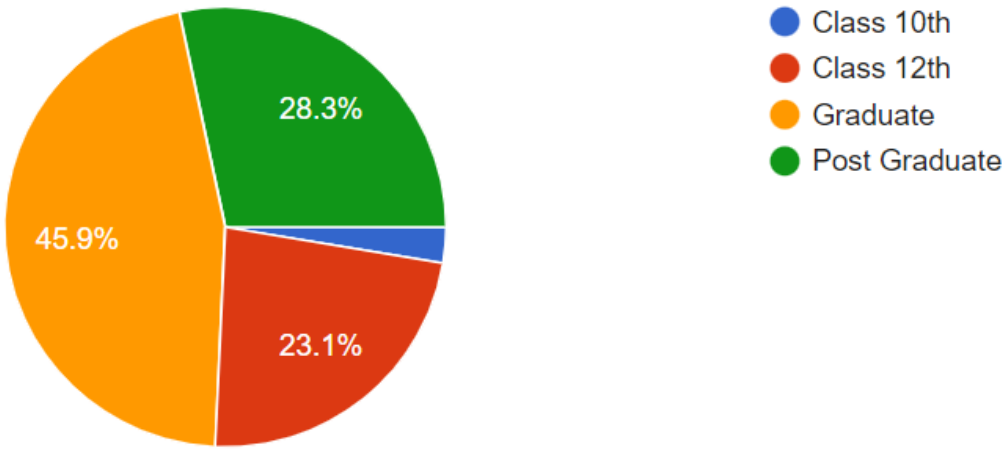

Figure 3.4: Educational qualification of the participant from Kerala.

3.3.2 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 is used to calculate a correlation between features, and the F-statistic can be used to calculate correlation with the class (relevance) (redundancy). The Pearson coefficient is given by,

$$r = \frac{\sum \frac{Xi-Xm}{Yi-Ym}}{\sum \frac{(Xi-Xm)^2}{(Yi-Y_mean)^2}} \tag{3.2}$$

where r is the Pearson coefficient,Xi is the values of x variable in a sample,Yi is the values of y variable in the sample,Xm is the mean of the values of the x variable,Ym is the mean of the values of the y variable.

Then, using a greedy search to maximize the objective function, which is a function of Relevance and redundancy, features are chosen.The phrase "Maximum Relevance - Minimum Redundancy" refers to the goal of selecting a feature at each iteration that has the greatest relevance to the objective variable and the least redundancy to the features chosen at earlier iterations. The MID and MIQ objective functions, which reflect the difference or quotient of relevance and redundancy, respectively, are two extensively utilized types of the objective function.

3.3.3 Boruta Algorithm

The Boruta algorithm wraps the Random Forest classification algorithm[14]. This approach claims to pick critical features by eliminating irrelevant attributes iteratively. By capturing all features that may occasionally be relevant to the result variable, Boruta uses an all-relevant feature selection methodology. Contrarily, the majority of conventional feature

selection algorithms use a minimal optimum approach, relying on a small collection of characteristics to produce a classifier with the lowest possible error.You can recursively remove features that performed poorly in each iteration of fitting a random forest model to a data set. As the strategy reduces the error of the random forest model, this will finally result in a minimal optimal subset of features. This occurs as a result of choosing an excessively pruned version of the input data set, which then discards some relevant features.The following are the steps of the Boruta algorithm:

Step I: Duplicates 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.

3.3.4 Synthetic Minority Over-sampling Technique (SMOTE)

The Synthetic Minority Oversampling Technique (SMOTE) is applied to address the class imbalance problem[14]. SMOTE uses feature space to generate synthetic samples of the minority class. The artificial models are introduced along the line 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 model. The multiplied result is then added to the feature vector in question, resulting in an artificial instance of the minority class.

3.4 Machine learning techniques for depression detection

To predict the existence of depression, the study has used five different machine learning classifiers, namely: KNN, Adaptive Boosting (AdaBoost), Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost) and Bagging classifier.

3.4.1 TabNet

A deep tabular data learning architecture called TabNet uses sequential attention to determine which features to infer from at each step of the decision-making process. A feature transformer, an attentive transformer, and feature masking make up the TabNet encoder. The processed representation is divided by a split block for the attentive transformer of the following phase as well as for the final output. The feature selection mask gives interpretable information about the functioning of the model for each step, and the masks can be combined to determine the importance of a feature globally. Each phase of the TabNet decoder consists of a feature transformer block.Fig 3.5 shows the overall architecture of TabNet.

3.4.2 K-Nearest Neighbor(KNN)

K-Nearest Neighbor is a simple Machine Learning algorithm that uses the Supervised Learning technique. The K-NN algorithm assumes similarity between the new case/data and

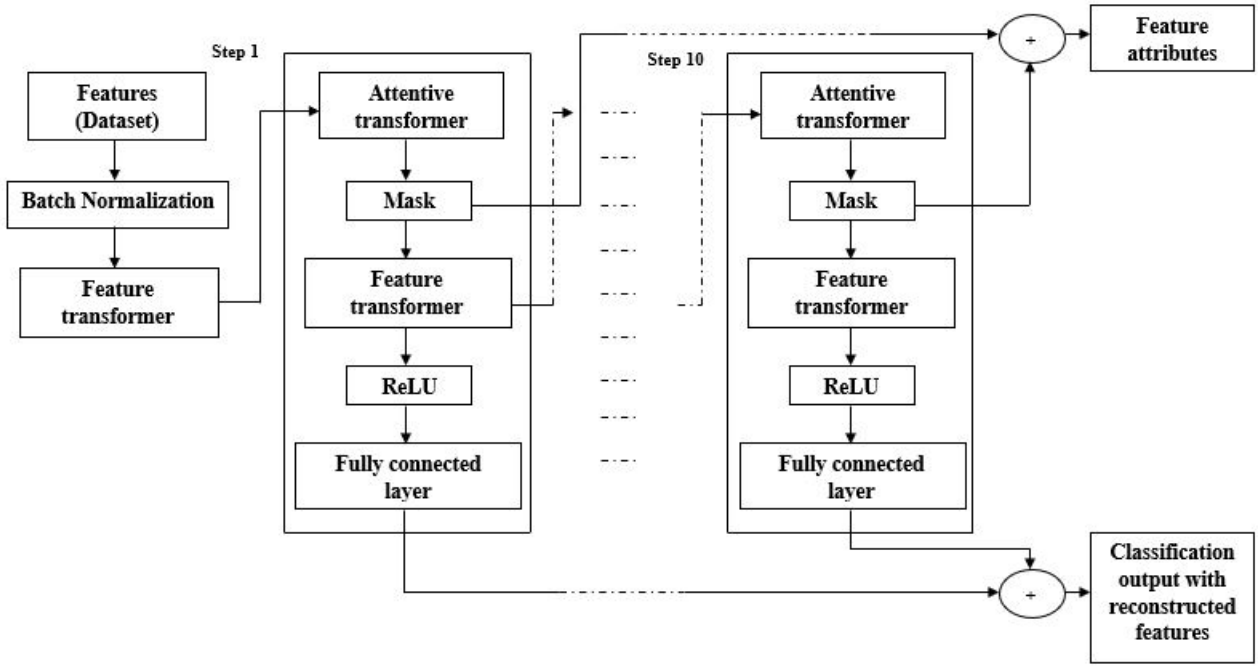

Figure 3.5: Architecture of TabNet.

existing cases and places the new case in the category that is most similar to the existing categories[5]. The K-NN algorithm stores all available data and uses similarity to classify new data points. This means that when new data is generated, it can be quickly classified into a well-suited category using the K- NN algorithm. The K-NN algorithm can be used for both regression and classification, but it is most commonly used for classification problems. K-NN is a non-parametric algorithm, which means it makes no assumptions about the underlying data.It is also known as a lazy learner algorithm because it does not immediately learn from the training set; instead, it stores the dataset and then performs an action on it during classification. During the training phase, the KNN algorithm simply stores the dataset and then classifies it into a category that is very similar to the new data.

3.4.3 Adaptive boosting (AdaBoost)

Boosting is an ensemble modelling methodology that seeks to produce a robust classifier from a large number of weak ones. It is accomplished by building a model out of a series of weak models. To begin, a model is built from the training data. The second model is then created, which attempts to correct the previous model's flaws. This procedure is repeated until either the entire training data set is adequately predicted or the maximum number of models is added. AdaBoost is a type of ensemble learning (also known as "meta-learning") that is designed to increase the efficiency of binary classifiers[2]. AdaBoost uses an iterative approach to improve poor classifiers by learning from their mistakes.

3.4.4 Gradient Boosting (GB)

Gradient Boosting (GB) classifiers combine a group of weak models to produce new models sequentially. 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[3]. A maximum number of models constructed or a threshold on predicted accuracy can be employed as a stopping criterion.

3.4.5 Extreme gradient boosting (XGBoost)

XGBoost is the abbreviation used fo r "Extreme Gradient Boosting". XGBoost is an extension to gradient boosted decision trees (GBM) and is specially designed to improve speed and performance. XGBoost is a distributed gradient boosting library developed to be very effective, adaptable, and portable. Under the Gradient Boosting framework, it implements machine learning techniques. It offers a parallel tree boosting to quickly and accurately address various data science challenges[4]. It has become approachable recently and is dominating applied machine learning and Kaggle competitions for structured data because of its scalability.

3.4.6 Bagging Classifier

Bagging was created using the ideas of bootstrapping and aggregation. Bootstrap datasets are made from the training dataset in the Bagging classifier. Each of these bootstrap datasets is utilized for training 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[2]. In many cases, combining the classifiers yields better results than using a single classifier. Because these characteristics are integrated into the Bagging classifier, it often outperforms other classifiers.

3.5 Implementation

The flowchart of the proposed model is shown in fig 3.6

3.5.1 Dataset splitting

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

3.5.2 Data encoding

After the completion of the Dataset Splitting technique, Data Encoding 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 Encoding which is an important pre-processing step for the structured dataset.

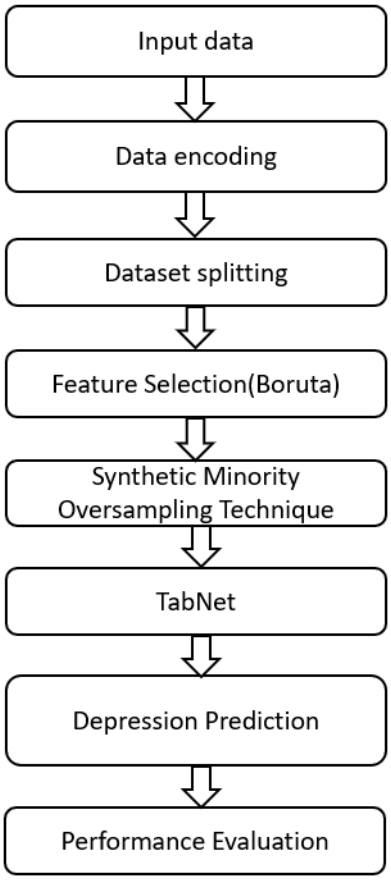

Figure 3.6: The flowchart of the proposed model

3.5.3 Modifying the training and testing dataset using feature selection

The presence of irrelevant features degrades the classifiers' efficiency. For extracting the relevant and necessary features from the dataset, three feature selection techniques have been used in this study separately.The feature selection techniques used were SelectKBest,Minimum redundancy and maximum relevance (mRMR)and Boruta Algorithm. 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 dataset collected from Bangladesh citizens and seven most relevant features have been selected from the dataset of Kerala citizens.

Table 3.2 shows the list of the selected features from dataset collected from Kerala citizens using these three feature selection techniques.Table 3.3 shows the list of selected features from dataset collected from Bangladesh citizens.Table 3.4 shows the list of selected features from combined dataset.

Feature Selection	**Selected features**	**Total**
SelectKBest	AGERNG, INFER, INSOM, ANXI, DEPRI, SUICIDE, CONFLICT, ABUSED, EATDIS, THREAT, POSSAT, WRKPRE, PROF, CHEAT, FINSTR	15
mRMR	ANXI, INSOM, INFER, PROF, DEPRI, SUICIDE, POSSAT, WRKPRE, CONFLICT, ENVSAT, TSSN, THREAT, EATDIS, FINSTR, GENDER	15
Boruta	AGERNG, PROF, POSSAT, INSOM, INFER, ANXI, DEPRI	7

Table 3.2: Selected features using feature selection techniques collected from Kerala citizens.

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

Table 3.3: Selected features using feature selection techniques collected from Bangladesh citizens.

Feature Selection	**Selected features**	**Total**
SelectKBest	INFER, DEPRI, ANXI, POSSAT, SUICIDE, INSOM, ABUSED, CONFLICT, CHEAT, FINSTR, THREAT, LOST, EATDIS, ENVSAT, WRKPRE	15
mRMR	ANXI, POSSAT, DEPRI, INFER, ENVSAT, INSOM, SUICIDE, FINSTR, CONFLICT, TSSN, ABUSED, LOST, THREAT, WRKPRE, CHEAT	15
Boruta	AGERNG, PROF, LIVWTH, ENVSAT, POSSAT, EATDIS, AVGSLP, INSOM, TSSN, WRKPRE, ANXI, DEPRI, ABUSED, CHEAT, THREAT, SUICIDE, INFER, CONFLICT, LOST	19

Table 3.4: Selected features using feature selection techniques in combined dataset.

3.5.4 Application of SMOTE on training data

Using an unbalanced dataset to train classifiers results in biased and erroneous predictions.Analysing the collected dataset from Kerala citizens it was found that the percentage of depressed and non-depressed subjects in the training datasets are 42.52% and 57.48%, respectively. SMOTE was utilized to solve the class imbalance problem in the training datasets because they were severely imbalanced. Table 3.4 shows the percentage of depressed and non-depressed participants before and after SMOTE in Kerala citizens.

State	Before SMOTE	After SMOTE
Depressed	42.52 %	100 %
Non-Depressed	57.48 %	100 %

Table 3.5: Depressed and non depressed before and after applying SMOTE in Kerala citizens.

Considering the available dataset from Bangladesh citizens,percentage of depressed and non-depressed subjects in the training datasets are 65.73% and 34.27%, respectively.Table 3.5 shows the percentage of depressed and non-depressed participants before and after SMOTE in Bangladesh citizens.

State	Before SMOTE	After SMOTE
Depressed	65.73%	100 %
Non-Depressed	34.27%	100%

Table 3.6: Depressed and non depressed before and after applying SMOTE in Bangladesh citizens.

In the combined dataset the percentage of depressed and non depressed before SMOTE were 79.98% and 20.02% respectively.Table 3.6 shows the percentage of depressed and non-depressed participants before and after SMOTE in the combined dataset.

3.5.5 Training and testing for predicting depression

This study used five different machine learning classifiers to predict the presence of depression: K-Nearest Neighbor (KNN), Adaptive Boosting (AdaBoost), Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), and Bagging classifier. Following training, each of these classifiers was used to predict the participants' depression.

State	Before SMOTE	After SMOTE
Depressed	79.98%	100 %
Non-Depressed	20.02%	100%

Table 3.7: Depressed and non depressed before and after applying SMOTE in combined dataset.

3.5.6 Performance evaluation

For all of the models built in the preceding sub-section, several performance measures like accuracy, precision, recall, and area under the curve (AUC) have been computed in this stage. These performance indicators are used to assess the efficacies of these models. Finally, based on the results of these performance criteria, the optimal model for predicting depression has been selected.

Chapter 4

RESULTS AND DISCUSSION

4.1 Results obtained using Dataset 1(Collected dataset from Kerala citizens)

Without using feature selection methods, the accuracies of the classifiers KNN, AdaBoost, GB, XGBoost and Bagging Classifier were 70.6%, 74.6%, 78.33%, 77.33%and 78% respectively. 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%. Using Boruta feature selection method on KNN classifier has yield the highest recall value of 92.3.

Fig 4.1 gives the comparative analysis of the accuracy using different feature selection techniques for the developed dataset from Kerala citizens .

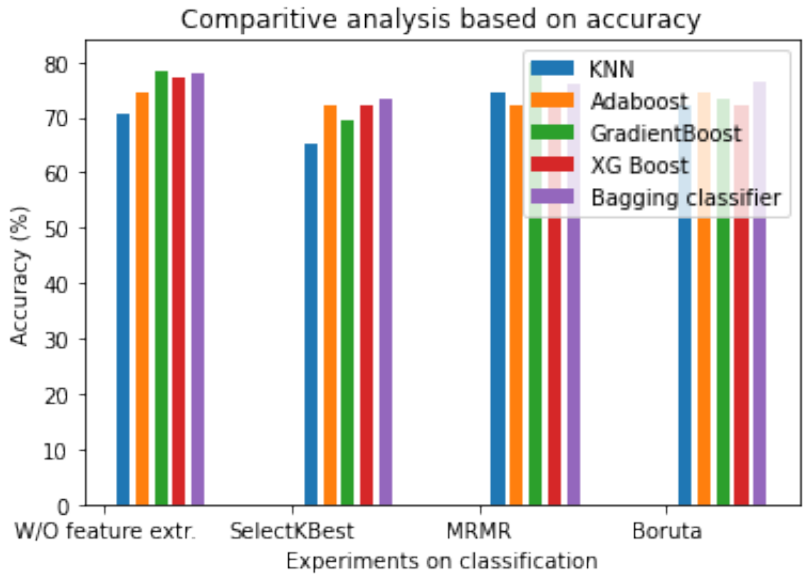

Figure 4.1: Comparative analysis of the accuracy of different feature selection methods for the developed dataset from Kerala citizens.

By applying the SelectKBest feature selection technique, the accuracies of the other classifiers namely, KNN, GB, XGBoost and Bagging are 65.33%, 72%, 69.3%, 72%, and 73.33% respectively.

In the case of using the MRMR feature selection technique, KNN, AdaBoost, GB, XGBoost and Bagging classifiers have attained accuracies of 74.6%, 72 %, 80%, 73.3 % and 76% 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%, 74.6%, 73.3%, 74%and 76% respectively.

These models' Area Under Curve (AUC) values have also been determined from the ROC Curve. If the AUC of a model is 1, it is regarded to be a perfect model or classifier. When a model's AUC is 0.5, it cannot differentiate between samples of various classes. As a result, a model with a higher AUC value is always preferred. The feature selection strategies improved classifier performance in terms of AUC, Precision, and Recall.

Sensitivity and specificity are important factors in evaluating a model's performance. A model with a higher sensitivity can identify participants with depression more accurately, while a model with a higher specificity can identify participants without depression more accurately.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 x axis representing the False Positive Rate and the y axis representing the True Positive Rate. Fig. 4.2 to 4.5 shows the ROC curves of these classifiers using different feature selection techniques applied on dataset collected from Kerala citizens. Fig. 4.2 to 4.5 reveals that the ROC curves of the classifiers have moved closer to the graph's upper left corner after applying feature selection techniques.

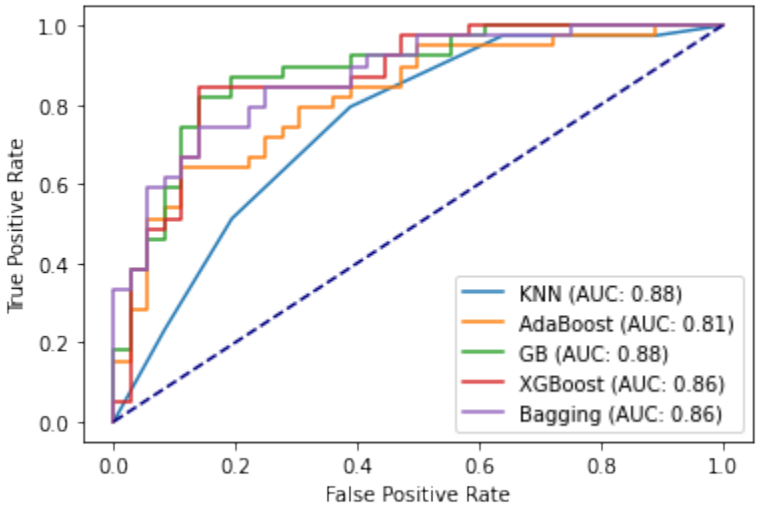

Figure 4.2: ROC Curve without using feature selection for the available dataset from Kerala citizens

Table 4.1 shows the confusion matrices of the predicted results for each of the classifiers using different feature selection techniques. Here,in the table the parameters used are explained below.

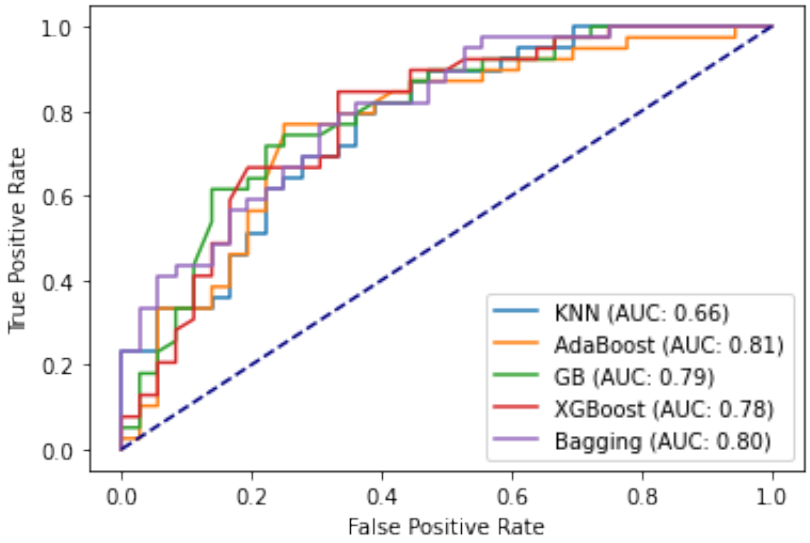

Figure 4.3: ROC Curve using SelectKBest for the available dataset from Kerala citizens

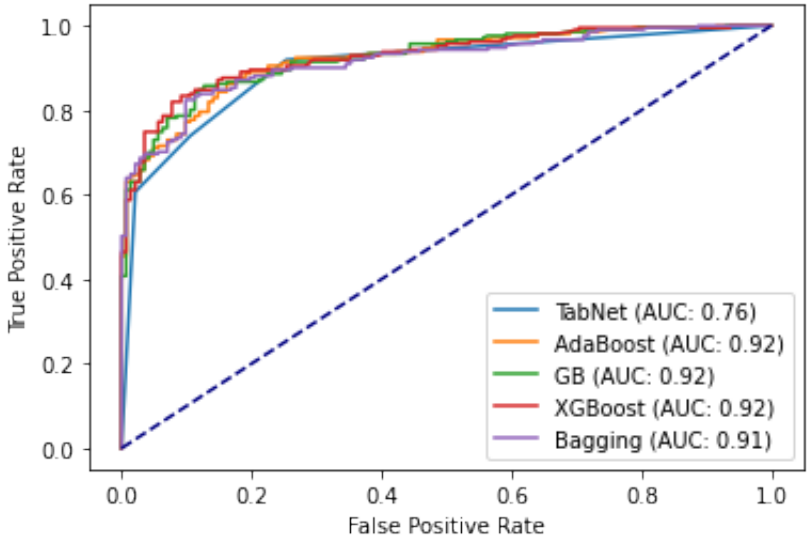

Figure 4.4: ROC Curve using mRMR for the available dataset from Kerala citizens

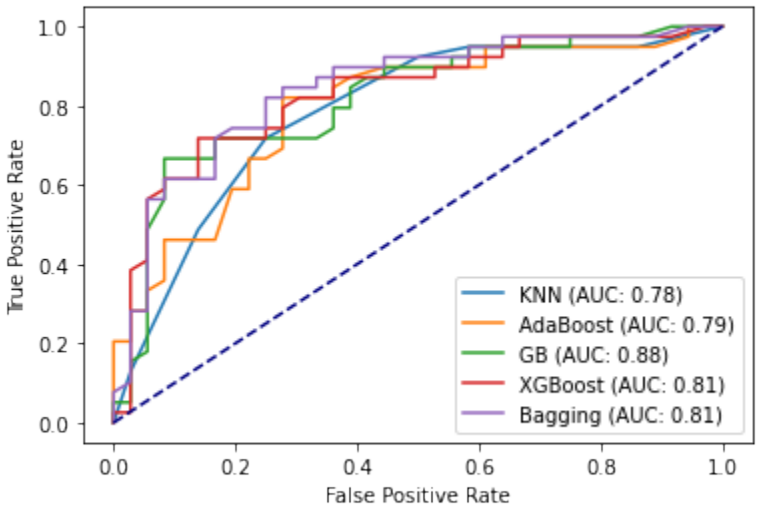

Figure 4.5: ROC Curve using Boruta for the available dataset from Kerala citizens

Model name	Feature Selection method	TP	TN	FP	FN
KNN	Without feature selection	31	22	14	8
AdaBoost		31	25	11	8
Gradient Boost		34	27	9	5
Extreme gradient boosting		34	27	9	5
Bagging classifier		33	27	9	6
KNN	Using SelectKBest	33	16	20	6
AdaBoost		32	22	14	7
GradientBoost		32	20	16	7
Extreme gradient boosting		29	25	11	10
Bagging classifier		32	23	13	7
KNN	Using mRMR	34	22	14	5
AdaBoost		34	22	14	5
GradientBoost		34	22	14	5
Extreme gradient boosting		31	24	12	5
Bagging		33	24	12	6
KNN	Using Boruta	36	18	18	3
AdaBoost		33	23	13	6
GradientBoost		34	21	15	5
Extreme gradient boosting		34	20	16	5
Bagging		34	23	13	5

Table 4.1: Confusion matrix obtained for the models using different feature selection methods on the available dataset from Kerala citizens

- True Positive (TP): When the classifier correctly predicts a participant is depressed, the result is known as True Positive (TP).
- True Negative (TN): When the classifier predicts that a participant without depression is not depressed, the result is True Negative (TN).
- False Positive (FP): When the classifier predicts a participant who is not depressed is depressed, the result is known as False Positive (FP).
- False Negative (FN): When the classifier predicts that a participant with depression is not depressed, the result is known as False Negative (FN).

Table 4.2 shows the Accuracy,Precision,Recall and AUC score obtained for different models using feature selection methods on the available dataset from Kerala citizens.

Model name	Feature Selection method	Accuracy	Precision	Recall	AUC
KNN	Without feature selection	70.6%	68.8%	79.4%	0.75
AdaBoost		74.6%	73.8%	79.48%	0.78
Gradient Boost		78.33%	79%	79.1%	0.86
Extreme gradient boosting		77.33%	79%	79.33%	0.86
Bagging		78%	78.5%	78.5%	0.86
KNN	Using SelectKBest	65.33%	62.26%	84.6%	0.66
AdaBoost		72%	69.56%	79.05%	0.76
Gradient Boost		69.33%	66.66%	75.05%	0.77
Extreme gradient boosting		72%	72.5%	74.35%	0.75
Bagging		73.33%	71.11%	76.05%	0.8
KNN	Using mRMR	74.66%	70.83%	87.17%	0.83
AdaBoost		72%	71.42%	76.92%	0.81
Gradient Boost		80%	77.27%	76.92%	0.81
Extreme gradient boosting		73.33%	72.09%	79.48%	0.80
Bagging		76%	73.33%	84.6%	0.83
KNN	Using Boruta	72%	66.66%	92.3%%	0.78
AdaBoost		74.66%	71.73%	84.61%	0.79
Gradient Boost		73.33%	69.38%	87.17%	0.81
Extreme gradient boosting		72%	68%	87.17%	0.82
Bagging		76.6%	72.34%	87.17%	0.84

Table 4.2: Accuracy,Precision,Recall and AUC score obtained for different models using feature selection methods on the available dataset from Kerala citizens.

According to Fig.4.6, various feature selection approaches greatly decreased the training time of the classifiers.Fig.4.6 indicates that feature selection strategies has increased the training time of the KNN classifier, while SelectKBest outperforms other feature selection techniques for minimising the training time of the AdaBoost,Boruta outperforms other feature selection methods for GB, XGBoost and Bagging classifiers.

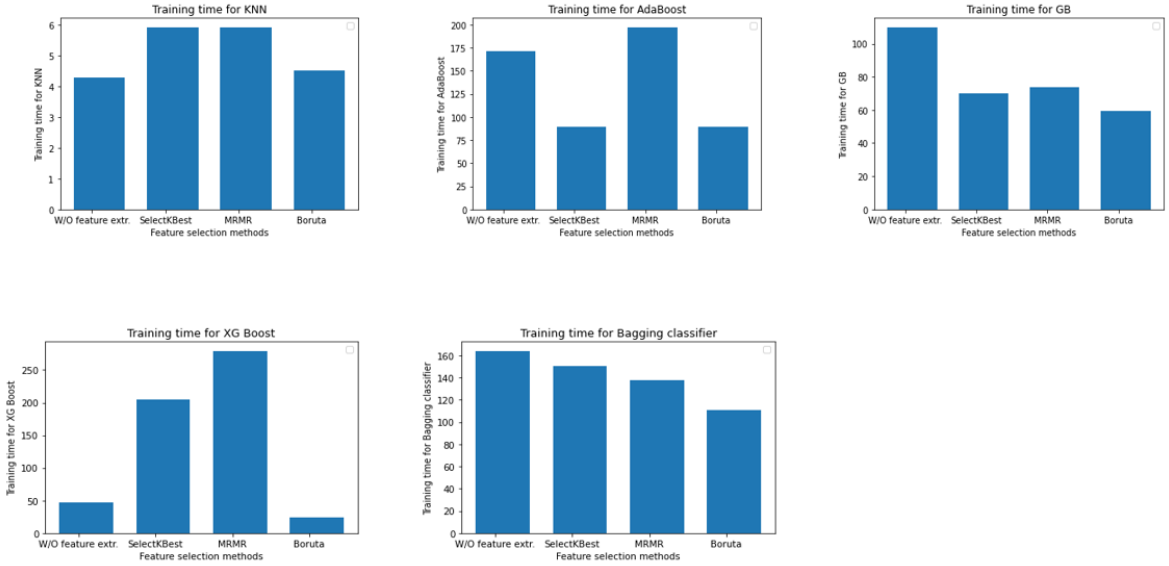

Figure 4.6: Training time of the classifiers for different feature selection techniques for the available dataset from Kerala citizens

4.2 Results obtained using Dataset 2(Available dataset from Bangladesh citizens)

Table 4.3 shows the confusion matrices of the predicted results for each of the classifiers using different feature selection techniques applied on dataset from Bangladesh citizens.

Without using feature selection methods, the accuracies of the classifiers KNN, AdaBoost, GB, XGBoost and Bagging Classifier were 66.94%, 87.60%, 86.78%,85.95%and 89.26% respectively. It can be concluded that the Bagging classifier produced the best results, with an accuracy of 89.26% without performing feature selection.

The AdaBoost outperformed the other classifiers in terms of accuracy when using the SelectKBest feature selection technique. It has a 92.56% accuracy rate. The accuracies of the other classifiers, namely KNN, GB, XGBoost and Bagging classifiers are 85.12%, 91.74 %, 86.78% and 90.91%respectively, when the SelectKBest feature selection technique is used.KNN, AdaBoost, GB, XGBoost and Bagging classifiers achieved 84.30 %, 91.74%, 90%, 90.08 % and 90.08% respectively, when using the mRMR feature selection technique.AdaBoost outperformed the other classifiers in terms of accuracy when using the Boruta feature selection technique. KNN, AdaBoost, GB, XGBoost, and Bagging classifiers achieved accuracies of 85.12%, 91.74%, 90.91%, 86.78%, and 90.08%, respectively.Fig 4.7 gives the comparitive analysis of the accuracy using different feature selection techniques.

Fig.4.8 to 4.11 shows the ROC curves of these classifiers using different feature selection techniques applied on the combined dataset.It can be seen that the ROC curves of the classifiers have moved closer to the graph's upper left corner after applying feature selection techniques.

The accuracy,precision,recall and AUC Score obtained for diferent models with and with-

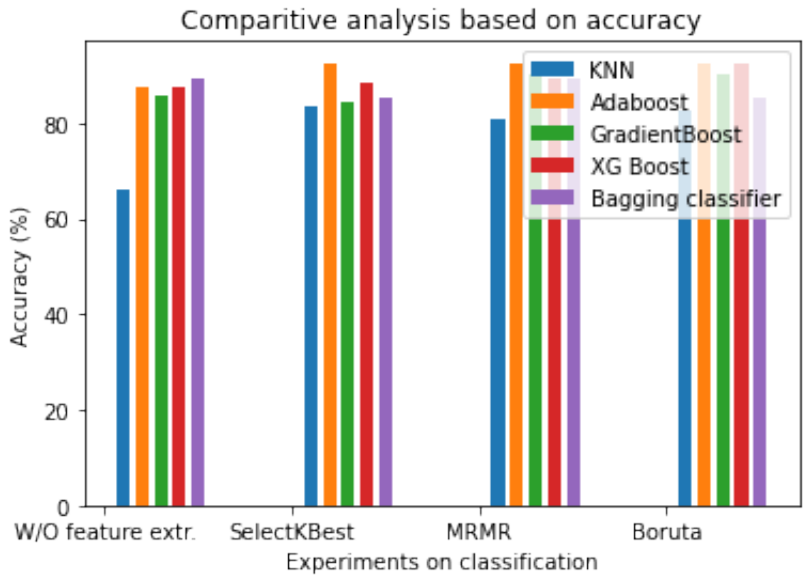

Figure 4.7: Comparitive analysis of the accuracy of different feature selection methods on dataset from Bangladesh citizens.

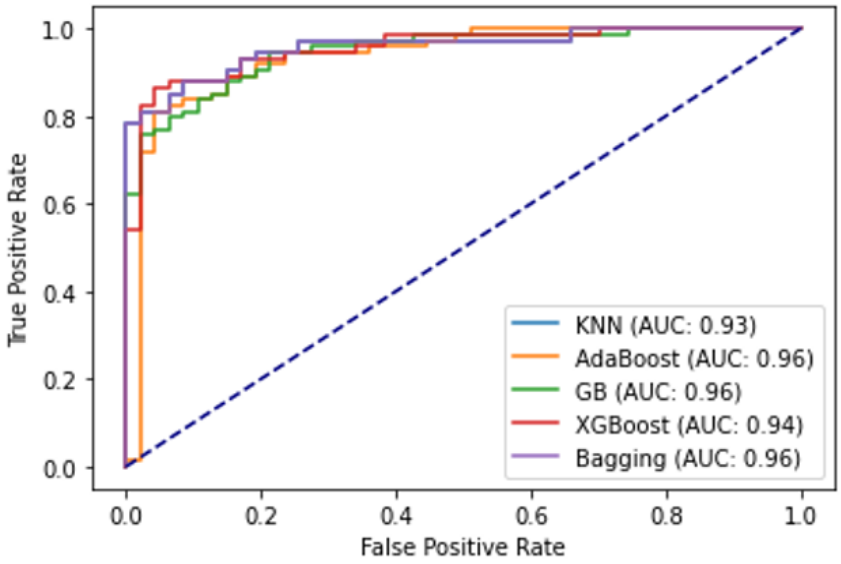

Figure 4.8: ROC Curve without using feature selection on dataset from Bangladesh citizens

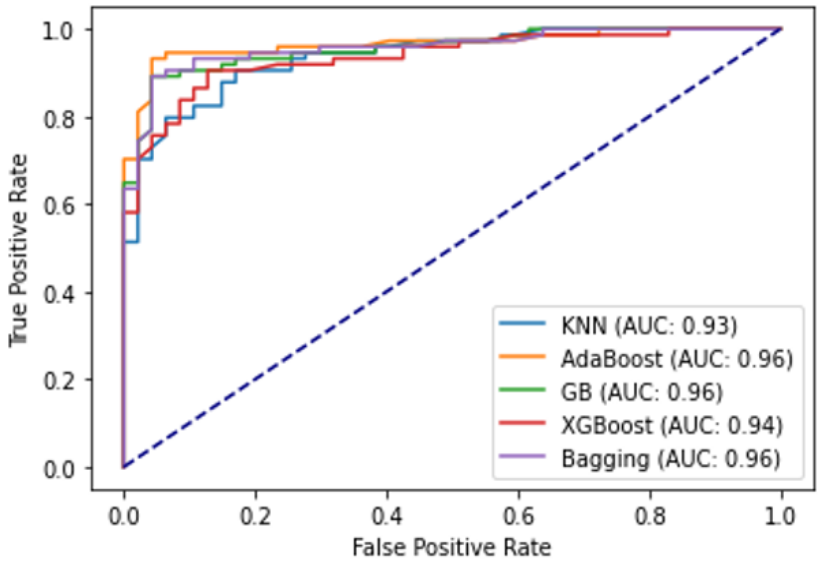

Figure 4.9: ROC Curve using SelectKBest on dataset from Bangladesh citizens

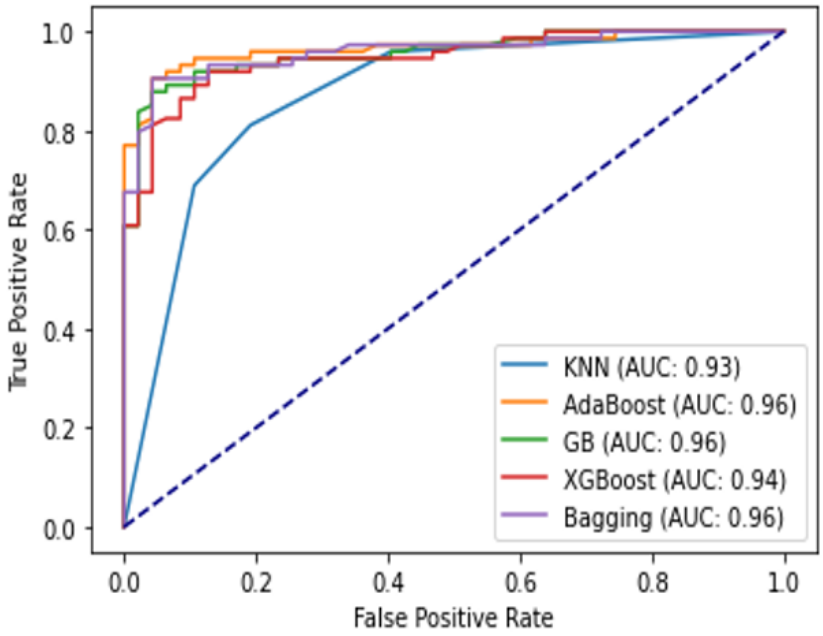

Figure 4.10: ROC Curve using MRMR on dataset from Bangladesh citizens

Model name	Feature Selection method	TP	TN	FP	FN
KNN	Without feature selection	40	41	6	34
AdaBoost		68	38	9	6
Gradient Boost		67	38	9	7
Extreme gradient boosting		67	37	10	7
Bagging classifier		69	39	8	5
KNN	Using SelectKBest	59	44	3	15
AdaBoost		68	44	3	6
GradientBoost		68	43	4	6
Extreme gradient boosting		67	38	9	7
Bagging classifier		67	43	4	7
KNN	Using mRMR	59	43	4	15
AdaBoost		67	44	3	7
GradientBoost		68	41	6	6
Extreme gradient boosting		68	41	6	6
Bagging		67	42	5	7
KNN	Using Boruta	61	42	5	13
AdaBoost		69	42	5	5
GradientBoost		68	42	5	6
Extreme gradient boosting		66	39	8	8
Bagging		68	41	6	6

Table 4.3: Confusion matrix obtained for the models using different feature selection methods on dataset from Bangladesh citizens

out feature selection methods are shown in Table 4.4.

The classifiers have become quicker as the number of features has been reduced utilising various feature selection strategies. According to Fig.4.12, various feature selection approaches greatly decreased the training time of the classifiers and it can be seen that mRMR outperforms other feature selection strategies for reducing the training time of the KNN classifier, while SelectKBest outperforms other feature selection techniques for minimising the training time of the AdaBoost, GB, and XGBoost classifiers. More than any other feature selection approach, the Boruta feature selection technique has reduced the training duration of Bagging classifiers. So it is found that the SelectKBest feature selection strategy outperforms other feature selection techniques in most cases in order to reduce classifier training time.

When the results of several models are compared, it can be determined that the AdaBoost classifier using the SelectKBest feature selection method outperformed the other models in terms of accuracy, AUC value, ROC-curves, and other measured performance metrics. The Adaboost classifier using the SelectKBest feature selection strategy achieved the greatest accuracy of 92.56%. This model also has the greatest AUC value of 0.96.

Based on the Maximum z-score, the Boruta feature selection method discards unimportant features.This method determined thirteen predictor variables to be significant. Based on the chi-score, the SelectKBest algorithm prioritised the predictor variables. The mRMR method, on the other hand, has ordered the predictor variables based on their relevance and redundancy. In this investigation, the top fifteen predictor variables were chosen from the rankings of both the SelectKBest and mRMR algorithms, because utilising the top fifteen

Model name	Feature Selection method	Accuracy	Precision	Recall	AUC
KNN	Without feature selection	66.94%	66.67%	55.4%	0.79
AdaBoost		87.60%	90.07%	91.89%	0.93
GB		86.78%	88.16%	89.18%	0.95
XGBoost		85.95%	87.01%	90.54%	0.96
Bagging		89.26%	89.61%	93.24%	0.96
KNN	Using SelectKBest	85.12%	95.16%	79.72%	0.93
AdaBoost		92.56%	95.77%	90.54%	0.96
GB		91.74%	94.44%	89.18%	0.96
XGBoost		86.78%	88.16%	90.54%	0.94
Bagging		90.91%	94.37%	90.54%	0.96
KNN	Using mRMR	84.30%	86.13%	81.08%	0.91
AdaBoost		91.74%	95.71%	90.54%	0.96
GB		90.08%	91.89%	91.89%	0.96
XGBoost		90.08%	91.89%	90.54%	0.95
Bagging		90.08%	93.06%	90.54%	0.96
KNN	Using Boruta	85.12%	92.42%	79.72%	0.91
AdaBoost		91.74%	93.24%	93.24%	0.96
GB		90.91%	93.15%	91.88%	0.96
XGBoost		86.78%	89.19%	91.89%	0.94
Bagging		90.08%	91.89%	91.89%	0.96

Table 4.4: Accuracy,Precision,Recall and AUC score obtained for different models using feature selection methods on the dataset from Bangladesh citizens.

predictor variables as input variables for the classifiers produces the best results in both situations.The results of the various models presented in this study confirm that the AdaBoost classifier with the SelectKBest feature selection technique is nearly the perfect model for predicting depression among the participants with an accuracy of 92.56%.

4.3 Results obtained using Dataset 3(Combined dataset)

Table 4.5 shows the confusion matrices of the predicted results for each of the classifiers using different feature selection techniques.

Without applying any feature selection techniques,the accuracies of the other classifiers namely, TabNet, GB, XGBoost and Bagging are 91.78%, 85.26%, 88.1%, 82.15%, and 82.15% respectively.

By applying the SelectKBest feature selection technique, the accuracies of the other classifiers namely, TabNet, GB, XGBoost and Bagging are 87.55%, 84.7%, 86.4%, 84.98%, and 84.98% respectively.

In the case of using the MRMR feature selection technique, TabNet, AdaBoost, GB, XGBoost and Bagging classifiers have attained accuracies of 90.36%, 85.26%, 86.4%, 84.98 % and 84.98% respectively.

In the case of using the Boruta feature selection technique,the achieved accuracies of TabNet, AdaBoost, GB, XGBoost and Bagging are 91.78%, 84.7%, 87.5%, 88.1%and 89.23% respectively.

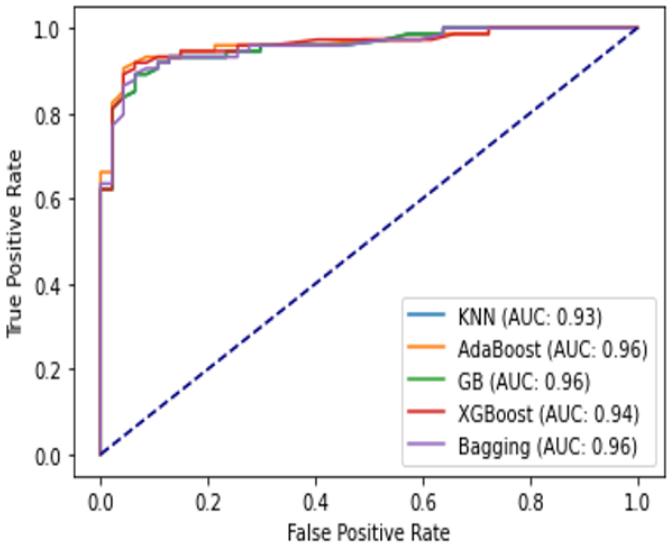

Figure 4.11: ROC Curve using Boruta on dataset from Bangladesh citizens

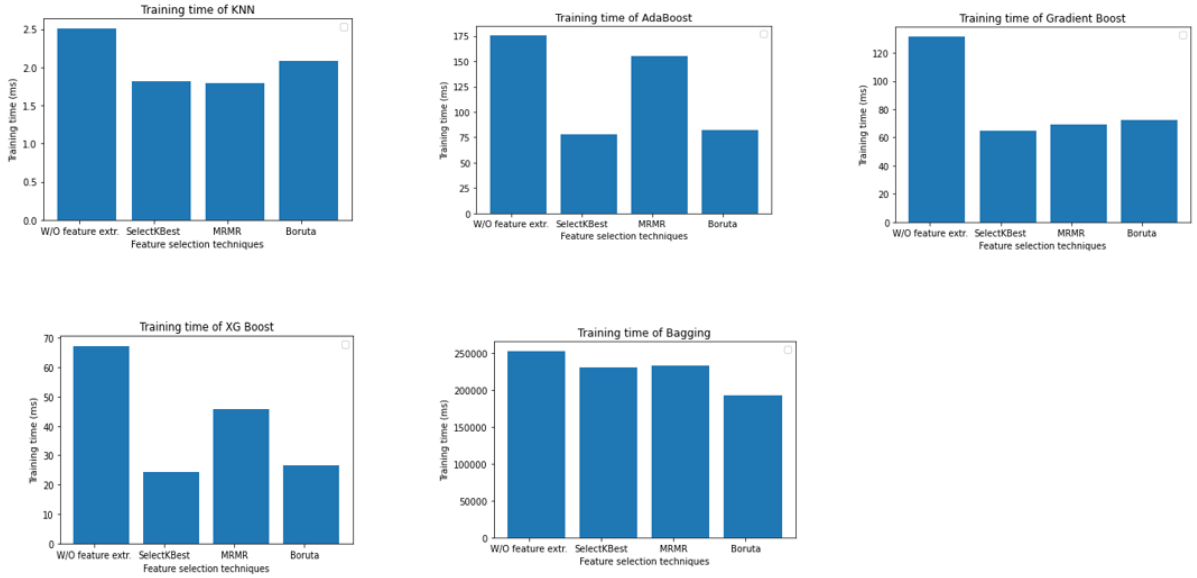

Figure 4.12: Training time of the classifiers for different feature selection techniques for the available dataset from Bangladesh citizens.

Model name	Feature Selection method	TP	TN	FP	FN
Tabnet	Without feature selection	194	130	17	12
AdaBoost		181	130	30	20
Gradient Boost		188	123	23	19
Extreme gradient boosting		188	123	23	19
Bagging classifier		184	127	27	15
TabNet	Using SelectKBest	178	124	33	18
AdaBoost		174	125	37	17
GradientBoost		178	127	33	15
Extreme gradient boosting		164	136	47	6
Bagging classifier		177	123	34	19
TabNet	Using mRMR	190	129	21	13
AdaBoost		156	127	55	15
GradientBoost		156	127	55	15
Extreme gradient boosting		184	121	27	21
Bagging		174	126	37	16
TabNet	Using Boruta	201	129	10	13
AdaBoost		181	118	30	24
GradientBoost		182	126	29	16
Extreme gradient boosting		185	126	26	16
Bagging		187	128	24	14

Table 4.5: Confusion matrix obtained for the models using different feature selection methods in combined dataset

Model name	Feature Selection method	Accuracy	Precision	Recall	AUC
TabNet	Without feature selection	91.78%	94.17%	91.94%	0.95
AdaBoost		85.26%	89.16%	85.75%	0.92
GB		88.1%	90.8%	89.09%	0.95
XGBoost		82.15%	82.45%	89.09%	0.9
Bagging		82.15%	92.46%	87.26%	0.95
TabNet	Using SelectKBest	87.55%	96.57%	84.36%	0.91
AdaBoost		84.7%	91.09%	82.46%	0.92
GB		86.4%	92.22%	84.36%	0.92
XGBoost		84.98%	96.47%	77.72%	0.92
Bagging		84.98%	90.3%	83.88%	0.93
TabNet	Using mRMR	90.36%	93.59%	90.04%	0.94
AdaBoost		85.26%	88.78%	86.25%	0.92
GB		86.4%	90.95%	85.78%	0.92
XGBoost		84.4%	89.75%	87.2%	0.92
Bagging		84.98%	91.57%	82.46%	0.91
TabNet	Using Boruta	91.78%	93.92%	95.26%	0.93
AdaBoost		84.7%	88.29%	85.71%	0.92
GB		87.5%	91.91%	86.25%	0.93
XGBoost		88.1%	92.03%	87.67%	0.94
Bagging		89.23%	93.03%	88.62%	0.94

Table 4.6: Accuracy,Precision,Recall and AUC score obtained for different models using feature selection methods on the combined dataset.

Table 4.6 shows the Accuracy,Precision,Recall and AUC score obtained for different models using feature selection methods on the combined dataset. Comparing the results of different models, it can be concluded that TabNet with and without using the feature selection algorithm has surpassed the other models in terms of accuracy,precision,recall and Area under the curve of ROC Curve values. TabNet classifier without using any feature selection technique has shown the highest accuracy,precision,recall and AUC values of 91.78%,94.17%,91.94% and 0.95 respectively.

Fig 4.13 gives the comparative analysis of the accuracy using different feature selection techniques for the combined dataset.From the comparative analysis of the accuracy it is clear that TabNet has shown the highest accuracy with and without using feature selection methods.Also it can be cocluded that TabNet has shown the maximum accuracy of 91.78% without using feature selection and also when Boruta feature selection was used.

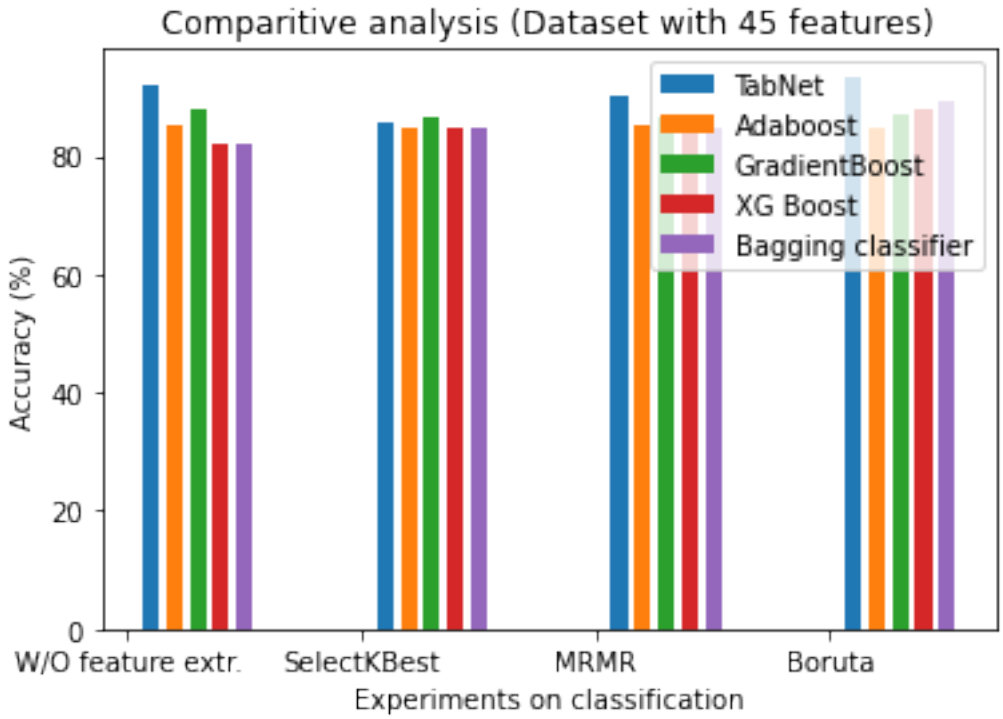

Figure 4.13: Comparitive analysis of the accuracy of different feature selection methods on combined dataset.

Fig.4.14 to 4.17 shows the ROC curves of these classifiers using different feature selection techniques applied on dataset from Bangladesh citizens.It can be seen that the ROC curves of the classifiers have moved closer to the graph's upper left corner after applying feature selection techniques.The AUC score can be interpreted from the Roc curve.When the AUC value approaches 1 it is assumed to be a perfect model.The AUC values obtained without using feature selection methods for TabNet, AdaBoost, GB, XGBoost and Bagging classifiers were 0.95,0.92,0.95,0.9 and 0.95 respectively.

The AUC values obtained using SelectKBest for TabNet, AdaBoost, GB, XGBoost and Bagging classifiers were 0.91, 0.92, 0.92, 0.92 and 0.93 respectively. The AUC values ob-

tained using mRMR for TabNet, AdaBoost, GB,XGBoost and Bagging classifiers were 0.94, 0.92, 0.92, 0.92and 0.91 respectively. The AUC values obtained using Boruta algorithm for TabNet, AdaBoost, GB, XGBoost and Bagging classifiers were 0.93, 0.92, 0.93, 0.94 and 0.94 respectively.

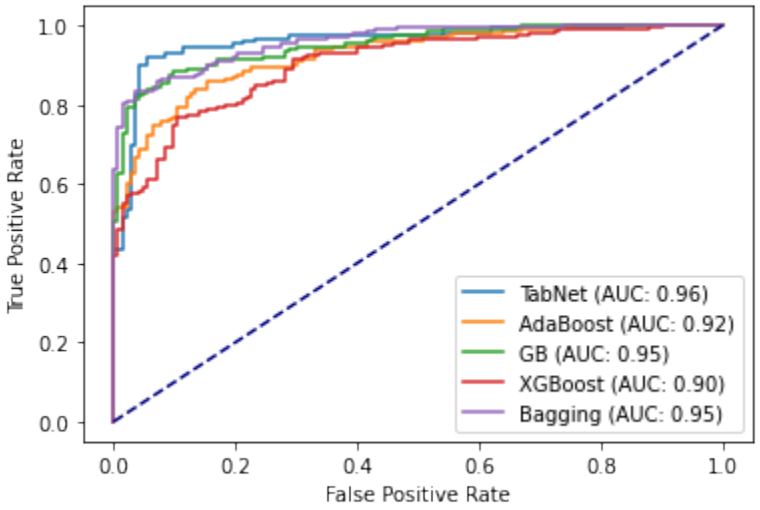

Figure 4.14: ROC Curve without using feature selection on combined dataset.

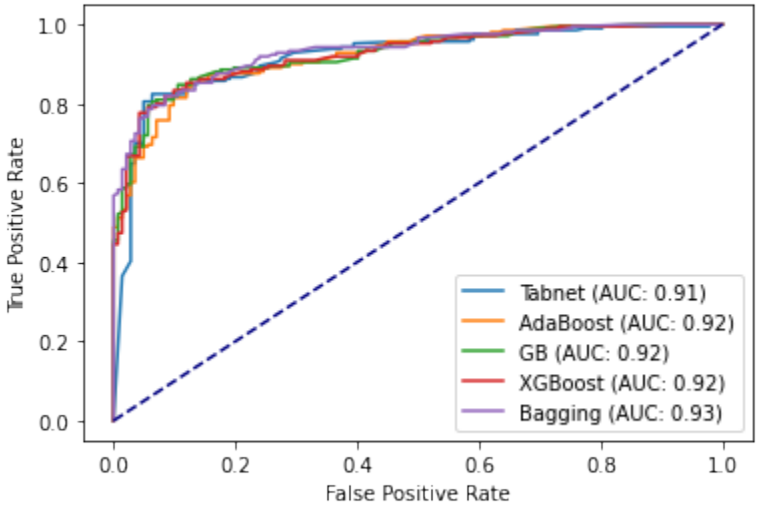

Figure 4.15: ROC Curve using SelectKBest on combined dataset

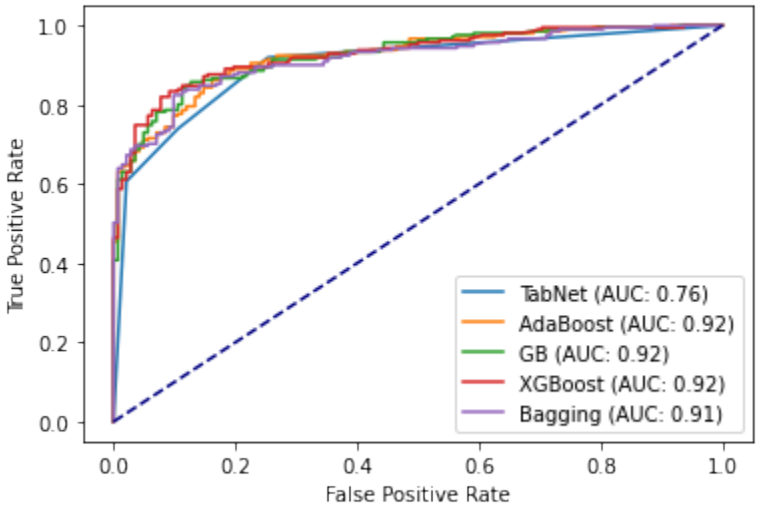

Figure 4.16: ROC Curve using MRMR on combined dataset

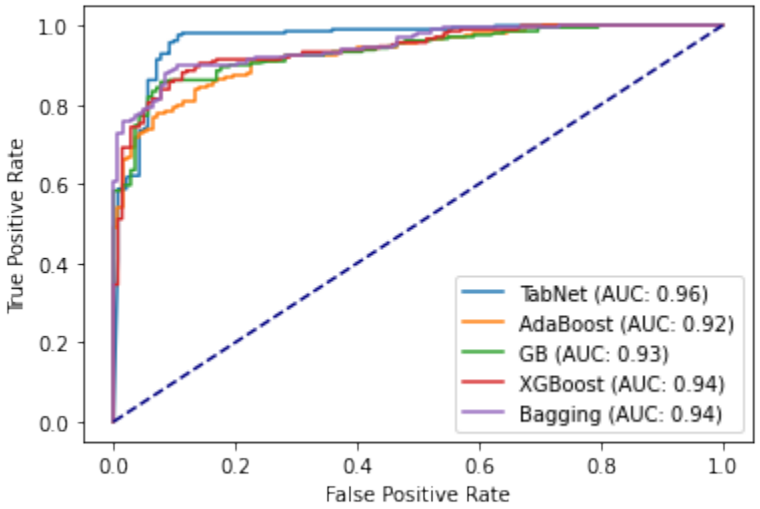

Figure 4.17: ROC Curve using Boruta on combined dataset

Chapter 5

CONCLUSION AND FUTURE WORKS

The major socio-demographic and psycho-social factors that contribute to the state of depression in a person are identified in this work. Initially, a dataset has been created, consisting of thirty socio-demographic, and psychosocial factors to screen for depression. Different feature selection techniques namely, SelectKBest, Minimum redundancy and Maximum relevance, and Boruta algorithm have extracted the most important demographic, and psychosocial factors responsible for forming a 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 sux different classifiers namely,TabNet, K Nearest Neighbor, AdaBoost, Gradient Boost, XG Boost, and Bagging classifier. By observing the outcomes of various models presented in this work, it can be confirmed that TabNet is the best model to predict depression among the participants with an accuracy of 91.78%.

This work has only made predictions about the presence of depression in people. In the future, this research can be expanded to determine a person's level of depression. Different biological aspects of the participants can be incorporated in the latter study as a result of the fact that a variety of biological elements have a notable impact on the development of depression in individuals. According to several studies, utilising various dimensionality reduction strategies during the data pre-processing processes enhances the performance of the models. These methods can be used, and the results of such applications can later be evaluated with those of the current study.

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 NP. Predicting anxiety, depression and stress in modern life using machine learning algorithms. Procedia Computer Science. 2020 Jan 1;167:1258-67.

[3] Sau A, Bhakta I. Screening of anxiety and depression among the seafarers using machine learning technology. Informatics in Medicine Unlocked. 2019 Jan 1;16:100149.

[4] Hatton CM, Paton LW, McMillan D, Cussens J, Gilbody S, Tiffin PA. Predicting persistent depressive symptoms in older adults: a machine learning approach to personalised mental healthcare. Journal of affective disorders. 2019 Mar 1;246:857-60.

[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] rotzek M, Koitka S, Friedrich CM. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Transactions on Knowledge and Data Engineering. 2018 Dec 18;32(3):588-601.

[7] Victor E, Aghajan ZM, Sewart AR, Christian R. Detecting depression using a framework combining deep multimodal neural networks with a purpose-built automated evaluation. Psychological assessment. 2019 Aug;31(8):1019.

[8] Shah FM, Ahmed F, Joy SK, Ahmed S, Sadek S, Shil R, Kabir MH. Early Depression Detection from Social Network Using Deep Learning Techniques. In2020 IEEE Region 10 Symposium (TENSYMP) 2020 Jun 5 (pp. 823-826). IEEE.

[9] 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.

[10] Uddin MZ, Dysthe KK, Følstad A, Brandtzaeg PB. Deep learning for prediction of depressive symptoms in a large textual dataset. Neural Computing and Applications. 2021 Aug 27:1-24.

[11] Mumtaz W, Qayyum A. A deep learning framework for automatic diagnosis of unipolar depression. International journal of medical informatics. 2019 Dec 1;132:103983.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

LIST OF PUBLICATIONS

Saji, G. V., Vazim, T., Sundar, S. (2021, November). Deep Learning Methods for Lung Cancer Detection, Classification and Prediction-A Review. In 2021 Fourth International Conference on Microelectronics, Signals Systems (ICMSS) (pp. 1-5). IEEE.