Array 18 (2023) 100291

Contents lists available at ScienceDirect

Array

journal homepage: www.elsevier.com/locate/array

AIDA: Artificial intelligence based depression assessment applied to Bangladeshi students

Rokeya Siddiqua, Nusrat Islam, Jarba Farnaz Bolaka, Riasat Khan, Sifat Momen *

Department of Electrical and Computer Engineering, North South University, Dhaka, 1229, Bangladesh

ARTICLE INFO

Keywords:
Depression assessment
Machine learning
Voting algorithm
Explainable AI

ABSTRACT

Depression is a common psychiatric disorder that is becoming more prevalent in developing countries like Bangladesh. Depression has been found to be prevalent among youths and influences a person's lifestyle and thought process. Unfortunately, due to the public and social stigma attached to this disease, the mental health issue of individuals are often overlooked. Early diagnosis of patients who may have depression often helps to provide effective treatment. This research aims to develop mechanisms to detect and predict depression levels and was applied to university students in Bangladesh. In this work, a questionnaire containing 106 questions has been constructed. The questions in the questionnaire are primarily of two kinds – (i) personal, and (ii) clinical. The questionnaire was distributed amongst Bangladeshi students and a total of 684 responses (aged between 19 and 35) were obtained. After appropriate consents from the participants, they were allowed to take the survey. After carefully scrutinizing the responses, 520 samples were taken into final consideration. A hybrid depression assessment scale was developed using a voting algorithm that employs eight well-known existing scales to assess the depression level of an individual. This hybrid scale was then applied to the collected samples that comprise personal information and questions from various familiar depression measuring scales. In addition, ten machine learning and two deep learning models were applied to predict the three classes of depression (normal, moderate and extreme). Five hyperparameter optimizers and nine feature selection methods were employed to improve the predictability. Accuracies of 98.08%, 94.23%, and 92.31% were obtained using Random Forest, Gradient Boosting, and CNN models, respectively. Random Forest accomplished the lowest false negatives and highest F Measure with its optimized hyperparameters. Finally, LIME, an explainable AI framework, was applied to interpret and retrace the prediction output of the machine learning models.

1. Introduction

Diagnosing depression is an extremely tricky and sensitive business in psychiatry. Depression, frequently referred to as the "common cold" of psychopathology, is the most well-known psychiatric condition, with a history that predates the study of psychiatry itself [1]. It is a complicated medical illness that affects patients to varying degrees. Depressed patients exhibit a variety of symptoms that may impair their ability to think, sleep, socialize, and even carry out routine tasks.

According to the World Health Organization (WHO), one in every four individuals suffers from mental health disorders worldwide [2], with over 280 million people of different ages suffering from depression [3]. In [4], it has been reported that mental diseases in Bangladeshis range from 6.5 to 31.0% among adults and 13.4 to 22.9% among children. A recent study claims that more than 7 million individuals in Bangladesh, predominantly urban and metropolitan areas, suffer from depression and anxiety [5]. Depression has been found to be highly prevalent among college and university students [6,7]. One survey reported that 24% of students suffer from both anxiety and depression [8]. The COVID-19 pandemic contributed to a further rise in the number of depression cases. According to one study, the COVID-19 pandemic has exacerbated depression among Bangladeshi students and has affected them in various ways, including increased sleep disorders and poor dietary habits [9]. The findings of these studies on Bangladeshi university students are concerning and require appropriate action. As a step toward that, this study attempts to take the first step towards understanding the primary causes of depression among the students studying at the tertiary level in Bangladesh and to build a mechanism to predict the severity of depression. A new scale for measuring depression levels has been developed, and machine learning techniques have been utilized to tackle this issue. Identifying depression at an early stage would undoubtedly prevent the situation from worsening further.

* Corresponding author.
E-mail address: sifat.momen@northsouth.edu (S. Momen).

https://doi.org/10.1016/j.array.2023.100291
Received 27 February 2023; Received in revised form 25 April 2023; Accepted 3 May 2023
Available online 10 May 2023

This study followed various measures to assess and predict depression, including establishing an integrated set of questionnaires and a new scale for measuring depression, collecting a private dataset, and analyzing and evaluating data. Following the creation of the questionnaires, data were collected in two ways: (1) via a Google form and (2) via a printed form. In both situations, a consent form was provided on the first page, and individuals were required to sign it to participate in this work. Major contributions of this work are illustrated below:

- In order to investigate depression levels amongst Bangladeshi university students, a dataset has been collected using a questionnaire in both online and offline format.
- A new scale combining eight well-known existing scales has been created to measure depression as either normal, moderate or extreme. A voting technique has been proposed to combine the eight existing scales.
- Machine learning and deep learning algorithms have been applied to classify three categories of depression. Hyperparameter optimizers and feature selection methods are used to improve the performance of these models.
- LIME, an explainable AI tool, has been used to understand the final predictions provided by machine learning models.
- The created models were cross-validated using an independently collected dataset from a different time period. The results demonstrate that the performance on the cross-validated dataset is comparable to the performance on the test dataset, demonstrating the robustness of the models.

To the best of our knowledge, this is the first time a new integrated scale based on a voting algorithm has been introduced, and a dataset of Bangladeshi participants has been used to investigate and comprehend the level of depression among university students.

The following is the order in which this article is organized. Related works are addressed in Section 2. Section 3 focuses discussion of the methodology adopted in this work. Section 4 examines the acquired outcomes. Section 5 discusses how the outcomes achieved our research objectives. Section 6 discusses the limitation of this work and finally, Section 7 includes a conclusion with remarks on our future work.

2. Related work

Psychiatric disorders, especially depression, are highly prevalent among young people with potentially severe complexities. Appropriate and early care for this disease is of paramount importance in dealing with its consequent impairing problems. However, it is not always possible due to associating social stigma and misconceptions with mental affliction treatment [10]. Considerable work has been performed to detect depression using measuring scaling systems and artificial intelligence techniques. This section briefly discusses some of the notable works in this area.

2.1. Related work on depression prediction using machine learning

In recent years, with the immense growth of artificial intelligence and machine learning techniques, many studies have been performed to detect depression and anxiety-related mental disorders employing a wide range of private and public datasets. Nemesure and his team utilized ensemble-based machine learning approaches to predict the depression and anxiety of undergraduate students [11]. More than four thousand students from the University of Nice Sophia-Antipolis participated in this study. The XGBoost model attained an AUC of 0.67 for the validation set. Lee and Kim applied machine learning frameworks to predict the depressive behavior of American adults [12]. They used the United States National Health and Nutrition Examination Survey (NHANES) dataset. NHANES conducted a PHQ-9 survey of approximately 8600 people for ten years. Boruta and LASSO feature selection approaches have been used in this work to select the dominant attributes. SVM, ANN and a wide range of ensemble models have been used to classify the depressed and non-depressed samples. SVM and ANN techniques achieved the highest accuracy and AUC of 77.1% and 0.813, respectively. Su et al. [13] proposed an automatic system to forecast the depression of the Chinese elderly population using the CLHLS survey dataset. KNN-based imputation framework has been used to substitute the missing data samples in the preprocessing step. Gradient-Boosted Decision Tree achieved the overall best performance with 75.9% accuracy and 0.63 AUC.

Ryu and colleagues developed a machine learning-based depression forecasting system for stroke survival patients [14]. NIHSS survey and Hamilton depression prediction index were performed on 623 individuals from a medical center in Korea. SVM and KNN reported accuracies of 77.5% and 73.3%, respectively. Haque et al. [15] developed an automated depression detection scheme for children using the Young Minds Matter dataset. The Boruta approach has been used to select the important features. The Decision Tree classifier produced 95% accuracy and 0.99 precision.

There has been considerable work on predicting depression among Bangladeshi individuals using locally collected datasets. For instance, Choudhury and his team used deep learning and five machine learning algorithms to predict depression among Bangladeshi undergraduate students based on some basic questionnaires [7]. Recursive Feature Elimination with Cross-Validation and Random Forest Classification was used for selecting salient features. A total of 65 questions were included in their dataset, including 7 basic information, 16 depression-related queries, 21 BDI and 21 DASS 21-BV (Bangla version) questions. After pre-processing, 577 participants were considered in this study. Zulfiker and his colleagues [16] used six different machine learning classifiers and three distinct feature selection methods to predict depression and extract relevant features. Synthetic Minority Oversampling Technique (SMOTE) [17] and Burns Depression Checklist (BDC) were also used in this work. The AdaBoost technique with the SelectKBest feature selection approach provided the best performance with 0.9256 classification accuracy.

Ahmed and teammates used five machine learning classifiers on two datasets to predict depression and anxiety [18]. In addition, two well-known depression and anxiety measuring scales were used in this work. The CNN model attained the highest accuracies of 0.96 for anxiety and 0.968 for depression detection employing 45 epochs. In [5], the authors utilized a binary logistic model to predict depression for Bangladeshi university students. The authors have collected a private dataset from an online survey of 210 participants using the DASS-21 scale. According to this study, relationships with parents and friends, bedtime and dietary patterns, and family socioeconomic status are the primary factors of depression and anxiety. Moon et al. employed various machine learning and ensemble approaches to predict the depression of employed Bangladeshis [19].

2.2. Related work on depression measuring scaling system

In order to measure depression levels, clinicians develop a set of questions that are given to an individual under assessment. The questions typically contain options that the individual needs to select. Based on the responses provided, a score is given per question. The total score obtained is then used to assess the level of depression of an individual. Different scaling systems exist, which are used by psychiatrists to diagnose the level of depression one is suffering [20].

A new integrated scale has been developed in this work using a majority voting technique on eight existing depression assessment scales, i.e., BDI, HDRS, MADRS, EQ-5, PHQ-9, QIDS-SR, DASS-21 and K-10. These assessment scales have been successfully utilized in many works to detect depression and anxiety-associated mental disorders. Some of these recent articles have been briefly presented below.

Table 1
A comparative analysis of related works.

Ref.	Subject	Sample size	Age range (years)	Data collection means	Advantage	Limitation
[11]	UG students	4184	18–20 mostly	Electronics health records	Investigates depression and anxiety disorder	Low AUC score
[15]	Children and adolescents	6310	4–17	Australian children and adolescent survey	The dataset contains rich feature set	The samples are in the age range 4–17
[13]	Adult chinese	1538	35–64	Chinese Longitudinal Healthy Longevity Study	The survey data was collected over a longer time period	Low accuracy
[14]	Stroke survivors	65	47–79	Comparable-aged stroke patients with 4-weeks screening were polled	Various cognitive and functional analysis were performed	Dataset with low sample size
[12]	Adults with hypertension	8628	≥40	PHQ-9 survey	The survey data was collected over a comparatively longer time period	Low accuracy
[7]	UG students	577	–	Beck Depression Scale and DASS-21	Multiple depression scales were used	Low accuracy
[16]	Bangladeshi participants	604	16 and above	Burns Depression Checklist (BDC)	Age distribution was wide	Depression level has not been predicted.
[18]	Bangladeshi women	–	15–35	Lifestyle related questions.	Multiple levels of depression and anxiety were measured.	Limited details on dataset distribution.

Ustun [21] used the Beck Depression Inventory scale to determine the depression level among the citizens of Turkey during COVID-19. A custom dataset with 1115 samples was collected through Google Forms over a span of ten days. Among the survey participants, 47%, 25.7%, 22.3%, and 5% were found to have minimal, mild, moderate, and severe depression, respectively. Saha and his colleagues [22] applied the Kessler K-10 metric to determine the level of psychological distress among Bangladeshi undergraduate students, particularly from Dhaka city. The authors collected 180 records using an online survey over a span of two weeks. Their dataset contained 28 features comprising coronavirus-related stress factors. This study discovered that almost 40% and 30.56% of students struggle with mild and moderate psychological distress, respectively. The Hamilton Scale for Anxiety (HAMA), Hamilton depression rating scale (HDRS), and Beck's Suicide Intent Scale were used in an observational study [23] with a set of predetermined questionnaires for Indian students. This study predicted anxiety, despair, and suicide intent among undergraduate dental students, as well as identified various stressors. Extended study hours, exhaustive workload, frequency of tests, competition among peers, and fear of failing were revealed to be statistically significant stressors.

Ibrahim et al. used Zagazig Depression Scale (ZDS) to measure the existence of psychological illness in a group of Egyptian undergraduate students [24]. In this work, ZDS, an individually-assessed Arabic language interpretation of the Hamilton Rating Scale was used to determine the pervasiveness of mental disorders symptoms. Participants revealed an average ZDS coefficient of approximately 18 and a maximum of 20. These ZDS scores indicate that more than 71% of the survey participants suffer from mild depression. Guo [25] used HAMD-17, CES-D, and WHOQOL-BREF evaluation metrics on undergraduate students with SSD to assess the impacts of electroacupuncture and cognitive behavioral therapy, or their combined effects, on mental disorders.

Ozawa conducted cross-sectional research on Japanese outpatients with ICD-10-defined depressive disorder [26]. Montgomery-Asberg Depression Rating Scale (MADRS) was used in this study with 100 depressed outpatients and 36 healthy family members. In [27], the authors directed exploration on a significant number of people with abrupt mood swings and subsyndromal depression symptoms. The integrated characteristics specifier was described as a score of 1 to 3 on selected items of the Montgomery Asberg Depression Rating Scale (MADRS) or Hamilton Depression Rating Scale (HAMD-17). According to Burchert and his colleagues [28], users of smartphone health applications are eager to analyze everyday depression symptoms. The Patient Health Questionnaire (PHQ-9) was used for analyzing short-term mood dynamics.

Wang and teammates [29] aimed to create a mapping framework that connects the acromegaly quality of life (AcroQoL) assessment survey to the EQ-5D-5L survey to provide a preference-based score that could be applied to study the socioeconomic assessment. For this study, 424 adult individuals had an average EQ-5D-5L coefficient score of approximately 0.80 with a 0.15 standard deviation. Mergen and colleagues evaluated the validity and accuracy of the Quick Inventory of Depressive Symptomatology (QIDS-SR16) and its American lexicon version using the Turkish student participants [30]. Lu et al. evaluated gender-based assessment invariance of the DASS-21 scale [31]. Among 13,208 students from five different cities, 4985 men and 8223 women participated in this study. The average indices for the DASS-21 depression, anxiety, and stress sub-scales were 2.17, 2.48, and 3.69, respectively.

Table 1 provides a summary comparison of related works. The following observations have been made after careful inspection of the related work – (i) there is an absence of using a hybrid scale to measure the level of depression, (ii) predictive models were typically applied on single and not on multi-scales, and (iii) most of the works have not produced an explanation of the prediction.

3. Methodology and model architecture

Fig. 1 shows the primary steps taken in this research. Following the creation of a questionnaire, it is distributed amongst Bangladeshi students and subsequently, a dataset is collected. The dataset is pre-processed so that it is in a format conducive to machine learning algorithms to operate on. The dataset is then organized into two forms – (i) dataset 1, which comprises 70 personal/basic questions of the respondents and (ii) dataset 2, which contains all the questions (106 questions), i.e., the personal questions as well as the clinical questions.

3.1. Questionnaires creation

One of the significant contributions of this work is to create a dataset of integrated depression assessment scales. The dataset is composed of two kinds of questions: (i) Personal questions and (ii) Clinical questions. After careful inspection of the literature, a total of nine

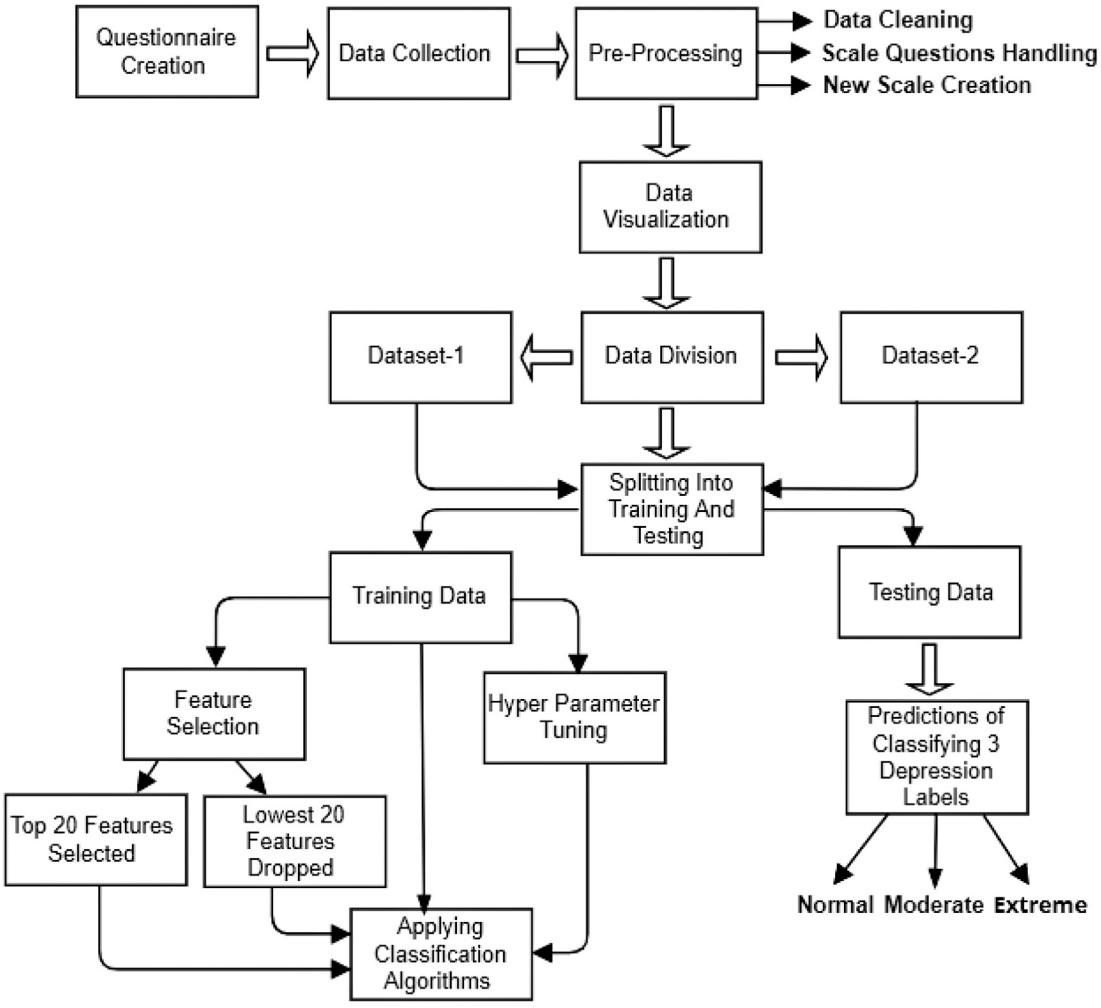


Fig. 1. Methodology of the work.

important areas were identified, which subsequently led to 70 personal/basic questions. Fig. 2 shows a mind map of the nine major areas and relevant parameters. An additional 36 questions have been created from various well-established depression assessment scales — BDI, HDRS, MADRS, EQ-5, PHQ-9, QIDS-SR, DASS-21 and K-10. Three random Bangladeshi undergraduate students (who had not seen those questionnaires before) were asked to take an initial survey to determine whether the selected questions were understandable to them. The initial survey took 20–25 min to answer all questions. The volunteers found several HDRS depression assessment scale questions confounding, including three questions from the personal questionnaire. Consequently, these HDRS questions were translated into the Bangla native language. Finally, the evaluation questionnaire was established.

3.2. Data collection

The survey was conducted online using Google Forms and offline (printed copy) over about nine and half weeks, i.e., from October 23, 2021, to December 28, 2021. A consent agreement document to participate in the survey was attached on the first page. Participants were given a brief idea about the study and their voluntary participation in the consent form. It was clearly stated that no traceable personal information would be collected and participants could stop taking the survey and withdraw at any time. A total of 684 records were collected — among those, 312 were offline and 372 were online submissions. 335 males and 349 females aged between 19 and 35 participated in the study.

3.3. Dataset preprocessing

The data cleaning, organizing, visualizing and finally, handling of the questions have been described exhaustively in this section.

3.3.1. Data cleaning

Some of the 684 collected records were incomplete, so they were removed from the final dataset. Only 148 of 312 offline submissions were considered after this removal process. The duplicate data checking and the feature-renaming process were conducted. Label and one-hot encodings were applied to provide a numerical representation of the categorical features. For example, ‘Undergraduate’ and ‘Postgraduate’ were set to 1 and 2, respectively, in the feature named ‘degree.’ Null-valued entries were replaced with their corresponding mean values. Min–max scaler was used to normalize features so that all numerical features have an acceptable range.

3.3.2. Selecting questions from familiar depression assessment tools

This work used eight depression measurement scales, i.e., BDI, HDRS, QIDS, MADRS, EQ5D, DASS21, PHQ-9 and K10. The selected questions from 5 scales (i.e., BDI, HDRS, QIDS, EQ5D, and DASS21) were taken verbatim. For the other three scales, if the questions overlap with the selected questions from the five scales, then they were not included — thus, we avoid duplication of questions.

To create each category, values were assigned to each question for that particular scale and scores were calculated by summing up

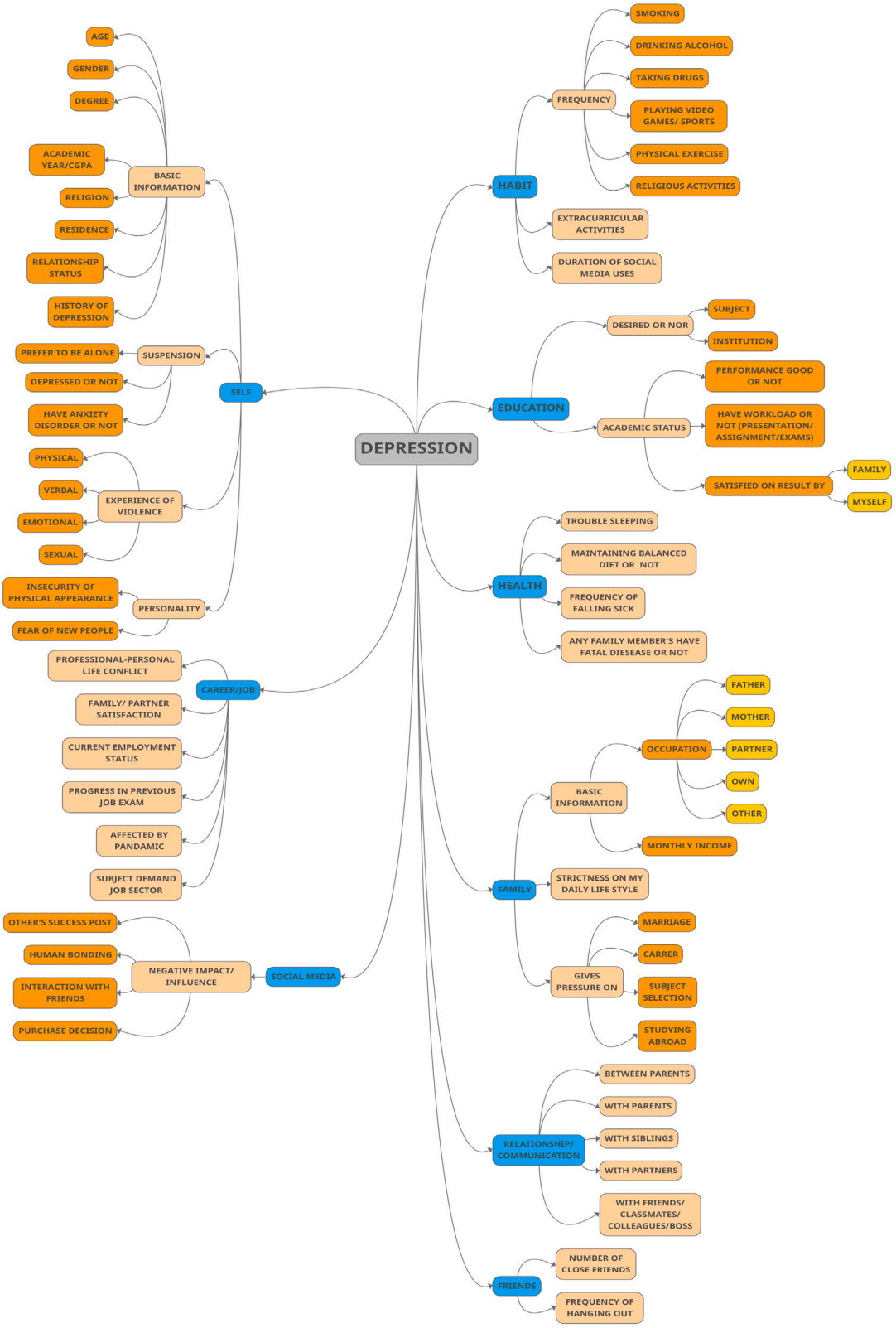


Fig. 2. Mind map of the personal questionnaires. The blue boxes show the nine major areas that were identified. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

those values. The scales have a unique range for measuring depression levels, primarily containing 4 to 6 classes (except EQ5D, which contains 2 classes). Finally, the range of the scales was normalized and expressed using three common depression classification levels, i.e., Normal, Moderate and Extreme, to maintain consistency and achieve better results.

Table 2
Synopsis of the eight depression measuring scales used in this study.

Scales	Levels of depression based on total score	No. of questions	Depression categories of new scale
Beck Depression Inventory (BDI)	Normal (1–10) Mild mood disturbance (11–16) Borderline clinical depression (17–20) Moderate depression (21–30) Severe depression (31–40) Extreme depression (over 40)	21	Normal (0–13) Moderate (14–27) Extreme (over 27)
Hamilton Depression Rating Scale (HDRS)	Normal (0–7) Mild depression (8–16) Moderate depression (17–23) Severe depression (over 23)	17	Normal (0–10) Moderate (11–22) Extreme (over 22)
The Quick Inventory of Depressive Symptomatology (QIDS-SR16)	Normal (0–5) Mild depression (6–10) Moderate depression (11–15) Severe depression (16–20) Very severe depression (21–27)	16	Normal (0–8) Moderate (9–17) Extreme (18–27)
Montgomery and Asberg Depression Rating Scale (MADRS)	Depressive symptoms absent (0–8) Mild (9–17) Moderate (18–34) Severe (35–60)	10	Normal (0–15) Moderate (16–30) Extreme (over 30)
EQ-5D	Worst (0) Best (100)	5	Normal (0–8) Extreme (over 8)
Depression, Anxiety and Stress Scale (DASS21)	Normal (0–9) Mild (10–13) Moderate (14–20) Severe (21–27) Extremely Severe (over 27)	21	Normal (0–13) Moderate (14–28) Extreme (over 28)
Patient Health Questionnaire (PHQ-9)	Minimal depression (1–4) Mild depression (5–9) Moderate depression (10–14) Moderately severe depression (15–19) Severe depression (20–27)	9	Normal (0–7) Moderate (8–16) Extreme (17–27)
Kessler Psychological Distress Scale (K10)	Normal (10–19) Mild disorder (20–24) Moderate disorder (25–29) Severe disorder (30–50)	10	Normal (0–22) Moderate (23–30) Extreme (31–50)

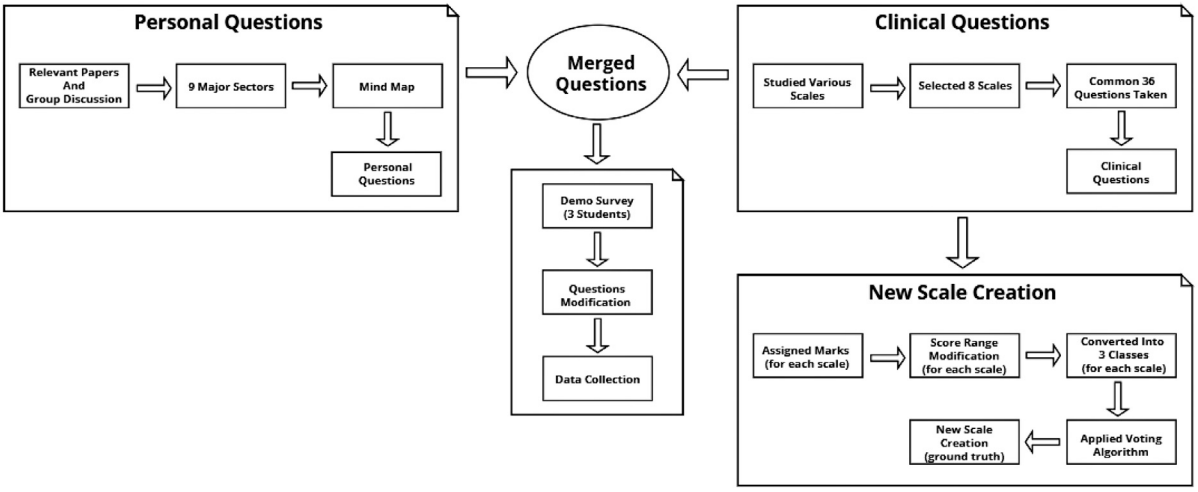


Fig. 3. Proposed depression assessment questionnaire and scale development procedure.

3.4. New depression assessment scale creation

In this research, a new scale was created using the voting technique on eight different depression measuring scales, discussed in Table 2. The eight scales do not use the same set of questions. One of the critical motivations for combining the eight scales is that it increases the feature space — this facilitates analyzing depression on a much wider number of factors. This also improves the predictability of the machine learning algorithms. To determine the level of depression for a new record, the new record is assessed against the eight depression scales. The most frequent label (normal, moderate, extreme) serves as the final depression level for this record. However, to handle equal

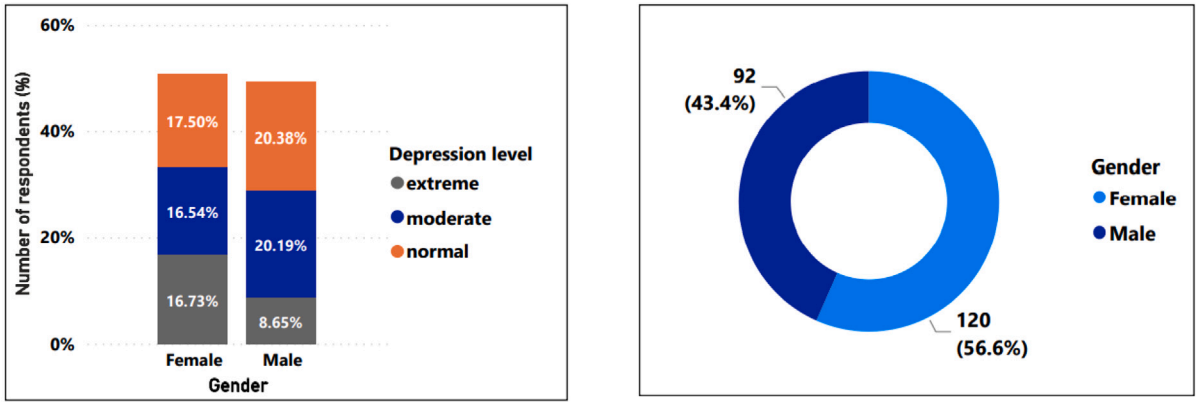


Fig. 4. Percentage distribution of respondent (a) depression level, (b) suicidal thoughts by gender.

votes, the lower depression level was taken into account. For example, if both 'normal' and 'moderate' levels obtain the maximum votes and are equal, then 'normal' was taken as the level of depression. Later, this new scale was used as a truth value while applying supervised machine learning techniques. The pseudocode of the voting algorithm is shown in algorithm 1. Finally, according to the new scale, 197 students were labeled normal, 191 students were marked moderately depressed, and 132 students were found extremely depressed. The depression detection questionnaire and the scale creation procedures are illustrated in Fig. 3.

Algorithm 1 Voting algorithm to measure depression

```
procedure VOTING ALGORITHM(Record r)
    scales = [BDI, HDRS, QIDS, MADRS, EQ5D, DASS21, PHQ-9, K10]
    normalFreq ← 0
    moderateFreq ← 0
    extremeFreq ← 0
    for i in scales do
        Determine depression level for Record r
        depLevel ← depression level for Record r
        if depLevel == normal then
            normalFreq ← normalFreq + 1
        else if depLevel == moderate then
            moderateFreq ← moderateFreq + 1
        else
            extremeFreq ← extremeFreq + 1
        end if
    end for
    if normalFreq ≥ moderateFreq and normalFreq ≥ extremeFreq then
        depressionLevel ← normal
    else if moderateFreq > normalFreq and moderateFreq ≥ extremeFreq then
        depressionLevel ← moderate
    else
        depressionLevel ← extreme
    end if
    return depressionLevel
end procedure
```

3.5. Data visualization

Fig. 4 provides information on the percentage of depression levels and suicidal thoughts among male and female students of the collected depression dataset. According to this figure, women are more vulnerable to both depression and suicidal thoughts than men. Depression was found in about one-third (33.27%) of the total respondents, whereas the male percentage was 28.84%. Fig. 4(b) shows the suicidal thoughts distributions across gender. Of the respondents who have suicidal thoughts, 56.6% (120 out of 212) were females and 43.4% (92 out of 212) were males.

Fig. 5 demonstrates the percentage of depression levels in those who have experienced emotional and sexual violence. According to Fig. 5, 71.23% and 37.73% of total participants admitted that they had faced emotional and sexual violence, respectively, which provokes depression in their life.

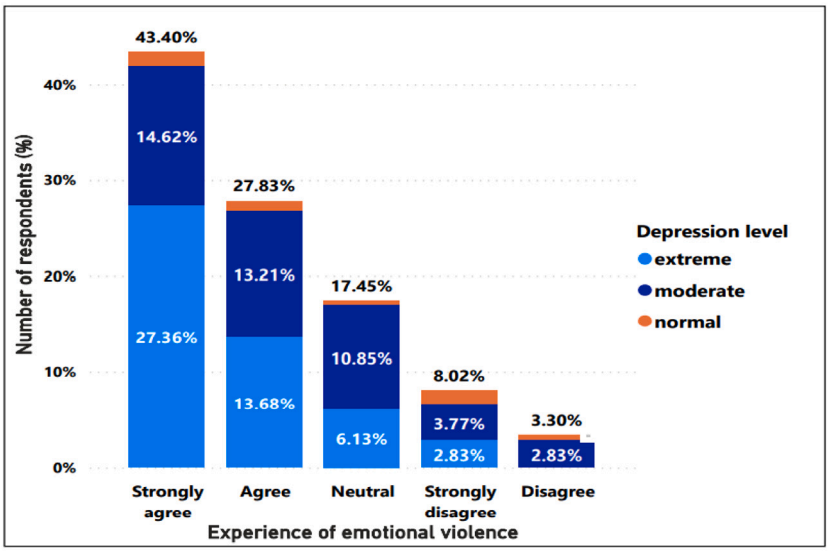
- Fig. 5(a) shows that both cases of "Strongly Agree" and "Agree" have a large number of extremely depressed students due to emotional violence where the normal percentage (1.42%) is very low. However, the Extreme case is almost twice (27.36%) than the moderate case (14.62%) according to the "Strongly agree" instances.
- Fig. 5(b) shows that a small portion of respondents have experienced sexual violence, but those who have been victims of sexual violence suffer from depression.
  According to Fig. 5(b), if observed carefully in both "Strongly Agree" and "Agree" instances, no student was found without depression. On top of that, the extreme depression class percentage is higher than the moderate percentage for sexual violence.

Fig. 6 shows the proportion of depression levels of survey respondents who have financial hardship. According to Fig. 6, more than half (57.41%) of students were found to have their families financially dependent on them due to their current financial difficulties and also have moderate to extreme level of depression.

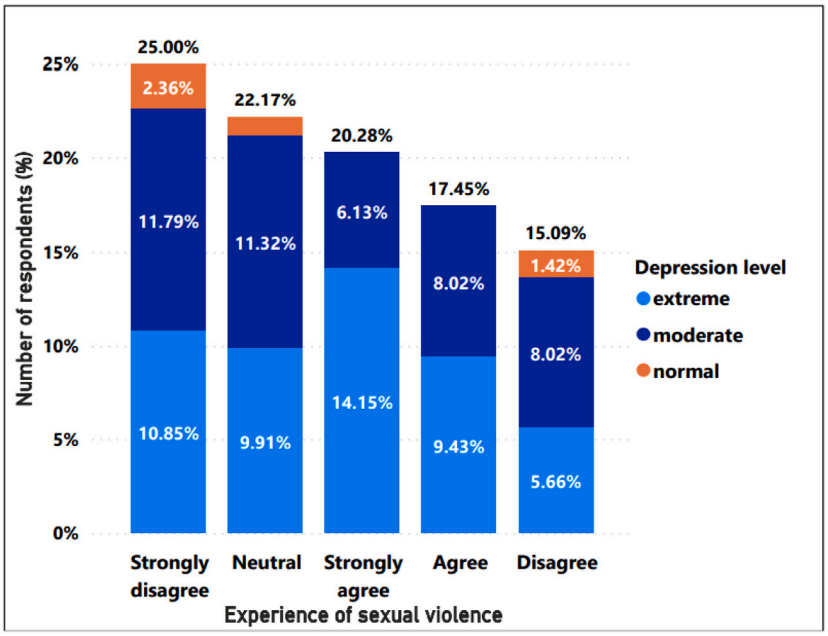
According to Fig. 7(a), 323 out of 520 (62.11%) of regular smokers have moderate to severe depression. It is also worth noting that there is a correlation between a lack of physical activity and depression. According to Fig. 7(b), 67.76% of participants who have never or infrequently engaged in physical exercise have moderate to severe depression.

Empirical findings also show that families play a critical role in people's lives, to the point where they feel pressured by family members. Participants were provided with six statements and asked to rate how much pressure they feel from their family members.

Fig. 8 indicates that the pressure towards career choice, higher studies and the lifestyle one adopt is slightly higher than the pressure one feels towards marriage. However, a significant amount of respondents feel extremely stressed with their role as a sibling or a child.



(a)



(b)

Fig. 5. Percentage of depression of survey respondents who have experienced (a) emotional violence and (b) sexual violence.

Since most of the participants are students, one source of depression originates from academic performance. Therefore, in order to understand the perception of the respondents in terms of their academic performances, seven statements were provided. The respondents used a 5-point Likert scale to specify their agreement level with these statements.

Fig. 9 shows that academic workload plays a critical role in the amount of stress they feel.

3.6. Algorithms application

After preprocessing the data, 520 records were obtained in the final dataset. The stratified approach was used to divide the dataset into two parts — thus ensuring that the two datasets contain equal representations from all classes. Dataset 1 contained only personal questions while dataset 2 included all questions, i.e., personal and questions obtained from the depression assessment scales. Both datasets were split for training (90%) and testing (10%) purposes. Therefore, 468 and 52 instances were assigned for the training and validation sets, respectively. Ten machine learning (Logistic Regression, Gradient Boosting, K-Nearest Neighbor, Random Forest, Decision Tree, Support Vector Machine, Perceptron, Naive Bayes (Gaussian), Naive Bayes (Multinomial), ZeroR Classifier and two deep learning (Artificial Neural Network (ANN), and Convolutional Neural Network (CNN)) techniques were applied.

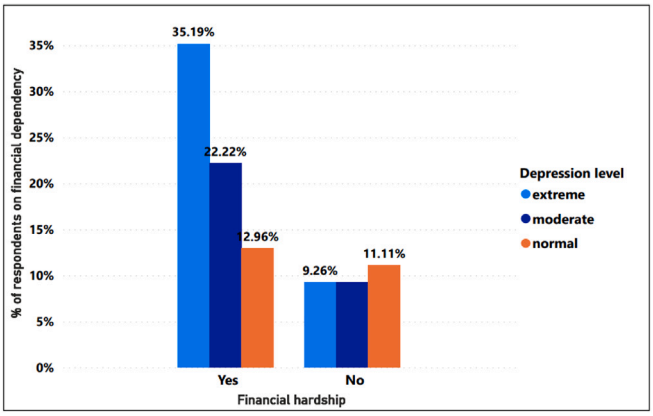


Fig. 6. Percentage distribution of depression level with financial hardship.

Logistic Regression uses the Linear Regression concept to solve classification problems. To solve the classification problem, it employs the sigmoid function (also known as the logistic function) on linear regression. Gradient Boosting employs an ensemble approach to combine many weak learners into a powerful predictive model. In contrast, K-Nearest Neighbor is an instance-based learning technique in which the prediction of a query record is based on the votes provided by the closest k training records. Random Forest uses an ensemble of Decision Trees to make predictions. Naïve Bayes algorithm uses a probabilistic approach to make the best prediction. Support Vector Machines are well known for their ability to solve classification and regression problems on both linear and non-linear data.

The Perceptron is a two-class (binary) classification algorithm. The output is predicted by calculating a weighted sum of the inputs, i.e., activation, and a bias (set to 1), expressed as:

$$f(x) = \begin{cases} 1, & \text{if } w \cdot x + b > 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

In Eq. (1), w, x and b denote weights, inputs and bias, respectively. $f(x)$ symbolizes the activation of the Perceptron model.

Gaussian Naive Bayes classifiers utilize Bayes' Theorem and Gaussian normal distribution to deduce the output. Interestingly, this advanced model exhibits an efficient feature, i.e., all the pairs of classified attributes do not rely upon each other. Consequently, the likelihood of the independent features is measured as follows:

$$P(x_j \mid y) = \frac{1}{\sqrt{2\pi s_y^2}} \exp\left(-\frac{(x_j - m_y)^2}{2s_y^2}\right) \tag{2}$$

where, m and s denote the mean and standard deviation of X. It is considered that, variance is independent of Y (i.e., s_i), or independent of X_i (i.e., s_k) or both (i.e., s). Multinomial Naive Bayes also follows essential steps, including transforming the samples of data into a set of frequencies, generating likelihood coefficients by obtaining the respective probabilities and calculating posterior probability using the Naive Bayesian equation as:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{3}$$

where, A and B denote discrete events, $P(A|B)$ is the probability of A given B is true and $P(B|A)$ signifies the probability of event B given A is true. $P(A)$ and $P(B)$ illustrate the independent probabilities of A and B, respectively. Finally, the output class is selected with the highest posterior probability. ZeroR is a useful classifier that helps to determine the baseline performance as a benchmark for other classification methods predicting the majority class.

Each layer of the Artificial Neural Network (ANN) is made up of multiple perceptron neurons. Since inputs are only processed forward, ANN is also known as a feed-forward neural network. In general, it has three distinct layers — input, hidden, and output layers and they are responsible for accepting inputs, processing them, and producing output, respectively. In this study, the Adam optimizer, mean squared error (as loss function), ReLU, and sigmoid activation functions were used to train over 15 epochs with a batch size of 5. Convolutional neural networks (CNN) are being used in a variety of applications and domains, particularly in the image and video processing works. In this study, the softmax activation function, Adam optimizer, categorical cross-entropy loss function, and two dense layers were used with a batch size of 5 to train over 15 epochs. Please see Fig. 10 for the architecture of the CNN model and Table 5 for its hyperparameters.

3.7. Hyperparameters optimization

In this work, five different hyperparameter optimization techniques (randomized search CV, grid search CV, hyperopt, TPOT classifier and Optuna) were used to find the best values for the hyperparameters for the proposed machine learning models. RandomizedSearchCV selects the optimal values by randomizing the search operation. In GridSearchCV, all values of hyperparameters are searched. The Bayes Theorem-based technique and a Python library, Hyperopt, are used in Bayesian Optimization. Natural selection, genetics concepts and a Python tool, Pipeline Optimization Tool (TPOT), are used in Genetic Algorithms. Optuna uses various optimization methods for finding the optimal hyperparameter values automatically.

3.8. Feature selection methods

In order to find out the salient features from both datasets, nine feature selection techniques were utilized, 1. Recursive Feature Elimination (RFE), 2. Univariate Selection (SelectKBest), 3. Fisher Score- Chi-squared Test, 4. Feature Importance (ExtraTreesClassifier), 5. Pearson Correlation, 6. Mutual Information, 7. Mutual Information Regression, 8. Manual Approach (Uniqueness) and 9. Variance Threshold. Recursive Feature Elimination selects the important features by removing the weakest feature individually until a specific number of features is obtained. SelectKBest class is used in the Univariate Selection method where the best features are selected based on the k highest score. Chi-squared Test approach provides highly dependent features which comprise a higher ChiSquare coefficient. The Chisquare is obtained by the following as:

$$X_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{4}$$

where, O and E denote the observed and expected count, respectively. The degree of freedom is expressed by C.

Gini Importance is used to select the features in ExtraTreesClassifier method and it selects the most significant k features. One of the two highly correlated features is dropped. The Pearson Correlation coefficient is given by:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5}$$

where, r denotes the correlation coefficient. Values of the x and y variables in a sample are given by x_i and y_i, respectively. $\bar{x}$ and $\bar{y}$ indicate mean of the x and y variables, respectively.

Mutual Information measures the mutual dependency between two random features. Mutual Info Regression also estimates mutual dependency for a continuous target variable. In the Uniqueness feature selection process, a significantly small number of unique value containing features were selected manually. In the Variance Threshold process, all features whose variance did not meet the threshold were removed.

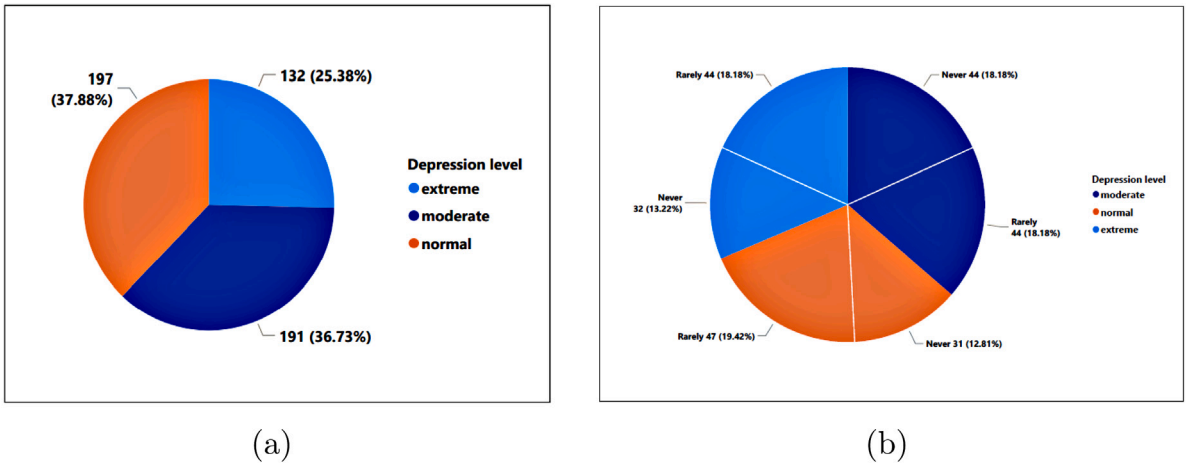


Fig. 7. Distribution of (a) smoking habit and (b) physical exercises of the respondents.

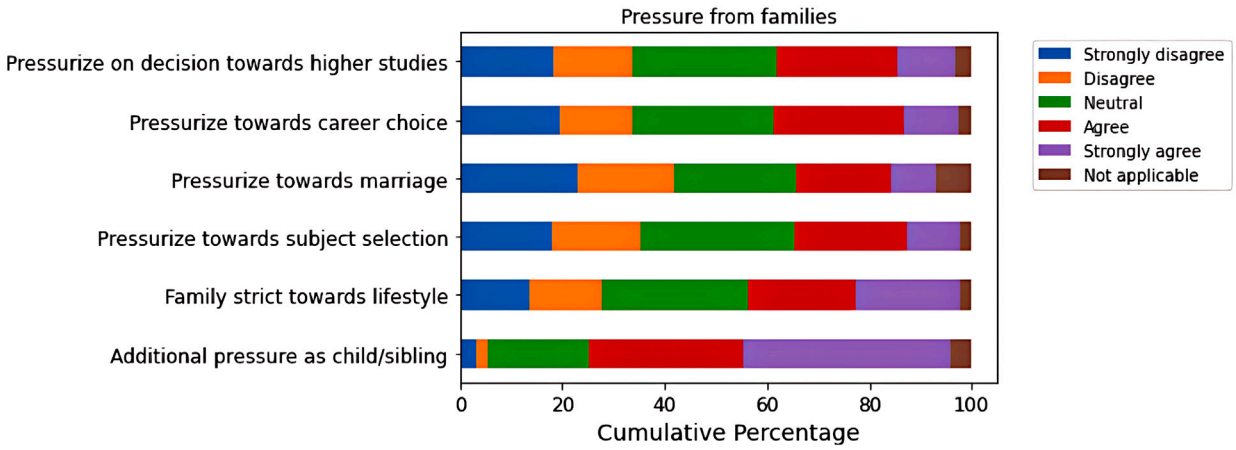


Fig. 8. Pressure from families.

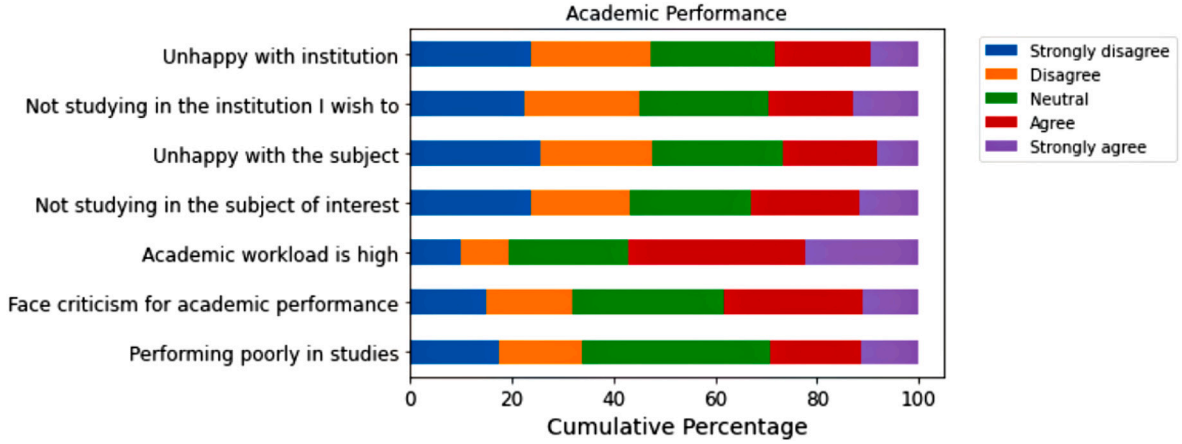


Fig. 9. Pressure from studies.

4. Result analysis

This section discusses the results of the proposed automatic depression assessment system.

Accuracy, precision, recall and F1 score are used to measure the performance of various machine learning and deep learning models. Eq. (6) to (9) were used to determine the performance metrics.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall} \tag{8}$$

Table 3
Performance metrics for different algorithms of dataset 1 and dataset 2.

Algorithms	Datasets	Accuracy	Precision	Recall	F1 score
Logistic Regression	Dataset-1	50.00%	48.79%	50.00%	48.97%
	Dataset-2	**88.46%**	88.50%	**88.46%**	88.42%
Gradient Boosted Algorithm	Dataset-1	50.00%	52.83%	50.00%	51.12%
	Dataset-2	82.69%	84.63%	82.69%	83.04%
K-Nearest Neighbor Algorithm	Dataset-1	**63.46%**	60.77%	61.54%	59.72%
	Dataset-2	78.85%	79.62%	79.62%	75.60%
Random Forest	Dataset-1	**63.46%**	62.67%	**63.46%**	**62.83%**
	Dataset-2	**90.38%**	89.17%	**88.46%**	**88.67%**
Decision Tree	Dataset-1	59.62%	63.14%	61.54%	61.44%
	Dataset-2	76.92%	77.54%	75.00%	75.45%
Support Vector Machine	Dataset-1	57.69%	58.17%	57.69%	57.82%
	Dataset-2	86.54%	**91.03%**	86.54%	86.53%
Perceptron	Dataset-1	46.15%	**65.21%**	32.69%	25.27%
	Dataset-2	76.92%	83.02%	75.00%	76.18%
Naive Bayes (Gaussian)	Dataset-1	55.77%	41.07%	55.77%	47.23%
	Dataset-2	73.08%	73.46%	73.08%	70.47%
Naive Bayes (Multinomial)	Dataset-1	61.54%	62.45%	61.54%	61.70%
	Dataset-2	86.54%	87.98%	86.54%	86.94%
ZeroR	Dataset-1	26.92%	07.24%	26.92%	11.42%
	Dataset-2	26.92%	07.24%	26.92%	11.42%
ANN	Dataset-1	48.08%	50.50%	38.46%	30.69%
	Dataset-2	69.23%	54.08%	65.38%	56.68%
CNN	Dataset-1	55.77%	58.08%	56.63%	57.72%
	Dataset-2	**92.31%**	88.83%	87.86%	88.63%

Table 4
Accuracy for different algorithms on dataset 1 and dataset 2 after using various hyperparameter optimizers.

Algorithms	Datasets	Randomized SearchCV (%)	GridSearch CV (%)	Bayesian Optimization (%)	Genetic Algorithms (%)	Optuna (%)
Logistic Regression	Dataset-1	48.08%	48.08%	48.08%	51.92%	50.00%
	Dataset-2	90.38%	**92.31%**	90.38%	84.61%	88.69%
Gradient Boosting	Dataset-1	**63.36%**	51.92%	55.77%	54.64%	53.85%
	Dataset-2	**94.23%**	79.90%	**94.23%**	89.20%	78.84%
K-Nearest Neighbor	Dataset-1	55.77%	55.77%	61.54%	55.77%	**63.46%**
	Dataset-2	73.08%	75.00%	**76.92%**	73.08%	73.08%
Random Forest	Dataset-1	59.61%	61.54%	**63.46%**	57.69%	61.53%
	Dataset-2	**92.31%**	**92.31%**	90.38%	92.31%	88.46%
Decision Tree	Dataset-1	55.77%	53.85%	53.85%	57.69%	**61.54%**
	Dataset-2	**92.31%**	69.23%	67.31%	69.23%	71.15%
Support Vector Machine	Dataset-1	57.69%	**64.75%**	58.90%	59.94%	57.90%
	Dataset-2	86.54%	86.54%	**94.23**%	86.54%	82.48%
Perceptron	Dataset-1	48.08%	48.64%	47.62%	49.64%	**53.94%**
	Dataset-2	75.00%	**84.62%**	83.61%	76.32%	80.69%
Gaussian Naive Bayes	Dataset-1	59.62%	59.62%	57.69%	55.77%	**61.54%**
	Dataset-2	78.85%	78.85%	**80.77%**	78.85%	78.85%
Multinomial Naive Bayes	Dataset-1	51.92%	53.85%	55.77%	**57.69%**	53.85%
	Dataset-2	**86.54%**	**86.54%**	84.61%	82.69%	84.61%
ZeroR	Dataset-1	26.92%	26.92%	26.92%	26.92%	26.92%
	Dataset-2	26.92%	26.92%	26.92%	26.92%	26.92%

Table 5
Hyperparameters of the CNN model.

Parameters	Values/Types
Epoch	15
Initial learning rate	0.001
Batch size	5
Optimizer	Adam
Loss function	Categorical cross-entropy

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

where, TP, TN, FP and FN are true positives, true negatives, false positives and false negatives respectively.

Table 3 shows the performance of ML and DL algorithms when applied to datasets 1 and 2. From Table 3 it is noticeable that KNN and Random Forest algorithms provide the highest accuracy of 63.46% when applied to dataset 1. Convolutional Neural Network (CNN) and Random Forest accomplished the accuracy of 92.31% and 90.38%,

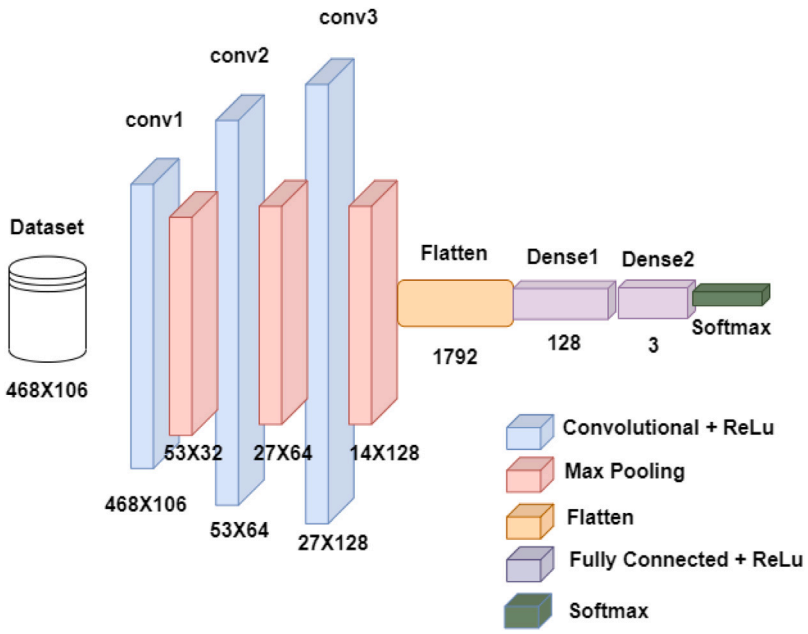


Fig. 10. Architecture of the CNN model.

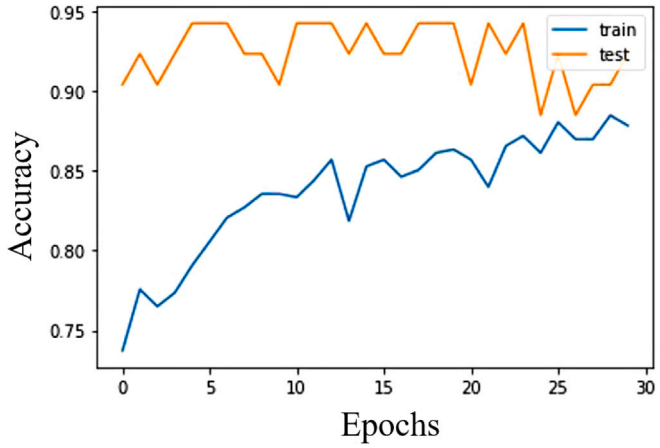


Fig. 11. Accuracy vs. epochs graph of CNN for dataset 2.

respectively, on dataset 2. The precision of 65.21% and 91.03% on datasets 1 and 2 respectively were achieved by Perceptron and Support Vector Machine. Recall of 63.46% and 88.46% were achieved on datasets 1 and 2 respectively using Random Forest. Logistic Regression also has 88.46% recall. As expected, the baseline classifier, ZeroR, achieved the lowest performance on datasets 1 and 2. The proposed ML models were found to provide better results on dataset 2.

Fig. 11 illustrates the accuracy with the change of epochs of the CNN model for dataset 2. As expected, the accuracy gradually improves with the increment of the epochs.

4.1. Performance of different algorithms with hyperparameter optimization

This section discusses the performance of various classifiers with hyperparameter optimization techniques applied to them. From Table 4, we can notice that after using RandomizedSearchCV, GridSearchCV and BayesianOptimization on dataset 1, 63.36%, 64.75% and 63.46% accuracies were obtained by Gradient Boosting, SVM and Random Forest frameworks, respectively. In the case of dataset 2, Gradient Boosting Algorithm and Random Forest achieved 94.23% and 92.31% accuracy with the RandomizedSearchCV optimizer.

As demonstrated in Tables 3 and 4, the performances of most of the depression assessment models improve significantly with the tuning of the hyperparameters. In dataset 2, the accuracy of the Gradient Boosting and Perceptron models increased from 82.69% to 94.23% and 76.92% to 84.62%, respectively, after optimizing hyperparameters. Similarly, in dataset 1, the accuracy of the Support Vector Machine increased from 57.69% to 64.75% after optimization. However, for Random Forest Classifier, the accuracy remains the same, which is 63.46% for both cases. SVM model with GridSearchCV and Gradient Boosting with RandomizedSearchCV attained the best accuracies for dataset 1 and dataset 2, respectively.

4.2. Performance analysis of different algorithms using various feature selection methods

The results obtained after taking the top 20 features and dropping the lowest 20 features from both datasets are shown in Tables 6 and 7 respectively. From Table 6, it can be seen that, on dataset 1, Gradient Boosting Algorithm and KNN provided 67.31% and 65.38% accuracies after using the Pearson Correlation and SelectKBest method respectively. On dataset 2, 92.31% and 96.15% accuracies were found from Logistic Regression and Random Forest using Pearson Correlation and Mutual Information methods respectively. Table 7 showed that on dataset 1, Gradient Boosting Algorithm provided 67.31% accuracy using the Pearson Correlation method. The Random Forest model with Pearson's Correlation coefficient-based feature selection technique attained the highest accuracy, i.e., 98.08%, for dataset 2.

Fig. 12 illustrates the comparison of accuracies between various ML techniques obtained before and after feature selection on both datasets with the most important 20 and 50 features. In dataset 1, the accuracy of Logistic Regression without feature selection is 50.00%, which increased to 63.46% and 65.38% after selecting the top 20 and 50 features, respectively. For Gradient Boosting Algorithm, the accuracy without feature selection is 50.00%. However, the accuracy improved to 67.31% considering both the best 20 and 50 features. Similarly, in dataset 2, Gradient Boosting Algorithm attained 82.69% accuracy, which increased to 90.38% and 88.54% after taking the dominant 20 and 86 features, respectively. The accuracy of Random Forest was 90.38% which increased to 96.15% and 98.08% after considering the most important 20 and 86 features, respectively.

Table 6
Accuracy for different algorithms on dataset 1 and dataset 2 after using various feature selection methods of significant 20 features.

Algorithms	Datasets	Recursive Feature Elimination (%)	Select K Best (%)	Fisher Score Chi square Test (%)	Extra Trees Classifier (%)	Pearson Correlation (%)	Mutual Information (%)	Mutual Info Regression (%)	Manual Uniqueness (%)	Variance Threshold (%)
Logistic Regression	Dataset-1	61.54	**63.46**	61.54	57.69	48.08	59.62	57.69	61.54	50.00
	Dataset-2	88.46	88.46	84.62	88.46	**92.31**	90.38	80.77	61.54	88.46
Gradient Boosting	Dataset-1	46.15	53.85	55.77	57.69	**67.31**	57.69	59.62	51.92	50.00
	Dataset-2	82.69	**90.38**	80.77	84.62	82.69	86.54	84.62	51.92	82.69
K-Nearest Neighbor	Dataset-1	51.92	**65.38**	**65.38**	57.69	57.69	59.62	61.54	55.77	63.46
	Dataset-2	**88.46**	82.69	82.69	76.92	76.92	76.92	78.85	55.77	78.85
Random Forest	Dataset-1	61.54	59.62	61.54	53.85	61.54	**65.38**	63.46	51.92	63.46
	Dataset-2	82.69	94.23	90.38	92.31	92.31	**96.15**	90.38	59.62	90.38
Decision Tree	Dataset-1	44.23	53.85	55.77	55.77	55.77	**63.46**	55.77	61.54	59.62
	Dataset-2	78.85	73.08	**90.08**	78.85	75.00	75.00	76.92	59.62	76.92
Support Vector Machine	Dataset-1	**63.46**	61.54	61.54	61.54	61.54	55.77	61.54	53.85	57.69
	Dataset-2	82.69	88.46	86.54	86.54	**90.38**	84.62	84.62	53.85	86.54
Perceptron	Dataset-1	46.15	**53.85**	34.62	48.08	34.62	36.54	51.92	44.23	46.15
	Dataset-2	**76.92**	53.85	**76.92**	**76.92**	65.38	67.31	71.15	44.23	**76.92**
Gaussian Naive Bayes	Dataset-1	53.85	**61.54**	59.62	59.62	55.77	53.85	55.77	61.54	55.77
	Dataset-2	78.85	86.54	80.77	**88.46**	73.08	82.69	82.69	61.54	73.08
Multinomial Naive Bayes	Dataset-1	**67.31**	57.69	61.54	55.77	61.54	61.54	51.92	59.62	61.54
	Dataset-2	80.77	82.69	82.69	84.62	**86.54**	78.85	**86.54**	59.62	**86.54**
ZeroR	Dataset-1	26.92	26.92	26.92	26.92	26.92	26.92	26.92	26.92	26.92
	Dataset-2	26.92	26.92	26.92	26.92	26.92	26.92	26.92	26.92	26.92

Table 7
Accuracy for different algorithms on dataset 1 (best 50 features) and dataset 2 (top 86 features) after using feature selection methods.

Algorithms	Datasets	Recursive Feature Elimination (%)	Select K Best (%)	Fisher Score Chi square Test (%)	Extra Trees Classifier (%)	Pearson Correlation (%)	Mutual Information (%)	Mutual Info Regression (%)	Manual Uniqueness (%)	Variance Threshold (%)
Logistic Regression	Dataset-1	55.77	48.08	50.00	44.23	48.08	53.85	57.69	**65.38**	50.00
	Dataset-2	**92.31**	90.38	90.38	88.46	**92.31**	88.46	88.46	88.46	88.46
Gradient Boosting	Dataset-1	55.77	50.00	50.00	50.00	**67.31**	55.77	53.85	65.38	50.00
	Dataset-2	82.69	78.85	78.85	78.85	82.69	82.69	84.62	**88.54**	82.69
K-Nearest Neighbor	Dataset-1	**63.46**	50.00	50.00	48.08	57.69	55.77	51.92	57.69	**63.46**
	Dataset-2	**80.77**	75.00	76.92	78.85	76.92	75.00	78.85	80.77	78.85
Random Forest	Dataset-1	61.54	**65.38**	57.69	**65.38**	59.62	57.69	55.77	63.46	63.46
	Dataset-2	92.31	86.54	88.46	88.46	**98.08**	86.54	88.46	86.54	90.38
Decision Tree	Dataset-1	59.62	57.69	61.54	61.54	53.85	**67.31**	57.69	57.69	59.62
	Dataset-2	67.31	69.23	73.08	73.08	73.08	73.08	75.00	67.31	**76.92**
SVM	Dataset-1	**63.46**	57.69	59.62	59.62	61.54	**63.46**	61.54	59.62	57.69
	Dataset-2	88.46	84.62	84.62	88.46	**90.38**	88.46	88.46	88.46	86.54
Perceptron	Dataset-1	48.08	**55.77**	46.15	53.85	34.62	32.69	28.85	42.31	46.15
	Dataset-2	**90.38**	76.92	**90.38**	67.31	65.38	80.77	63.46	73.08	76.92
Gaussian Naive Bayes	Dataset-1	53.85	55.77	55.77	57.69	55.77	**59.62**	57.69	57.69	55.77
	Dataset-2	78.85	76.92	73.08	76.92	73.08	75.00	75.00	**86.54**	73.08
Multinomial Naive Bayes	Dataset-1	63.46	59.62	**65.38**	59.62	61.54	61.54	55.77	**65.38**	61.54
	Dataset-2	88.46	84.62	84.62	86.54	86.54	**90.38**	86.54	86.54	86.54
ZeroR	Dataset-1	26.92	26.92	26.92	26.92	26.92	26.92	26.92	26.92	26.92
	Dataset-2	26.92	26.92	26.92	26.92	26.92	26.92	26.92	26.92	26.92

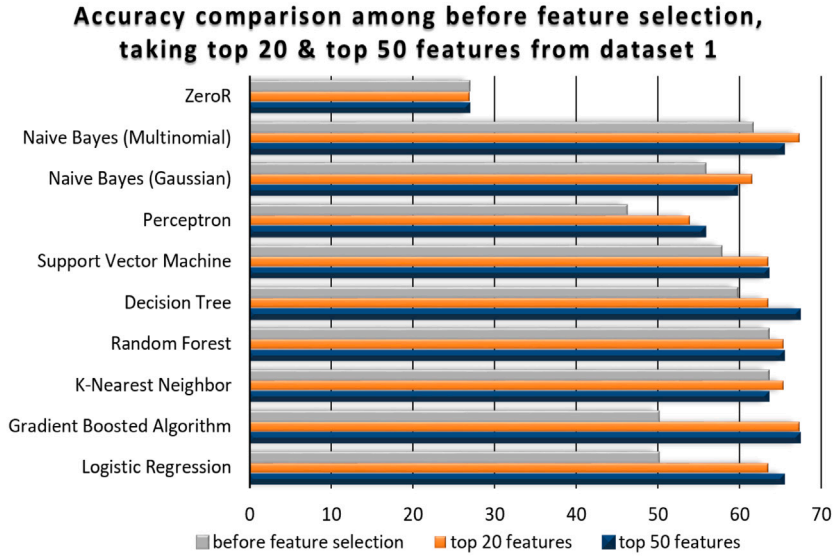
Synopsis of the feature selection techniques for dataset 1 and dataset 2 are summarized in Table 8.

Local Interpretable Model-Agnostic Explanations (LIME) is an efficient technique for comprehending black box machine learning models [32]. This technique constructs a simpler surrogate model by locally approximating the complex ML model. The surrogate model analyzes the confined area of the individual prediction and formulates a logical explanation in that local region. Figs. 13 and 14 demonstrate the depression prediction interpretation of an extreme case and a normal instance, respectively, provided by the LIME explainable AI framework. The Random Forest model with the Pearson correlation feature selection technique predicted extreme depression with a 0.88 confidence score for this specific sample, as illustrated in Fig. 13, because of its irregular sleep pattern, suicidal and somatic symptoms, disappointment with the future and self-blaming attitude.

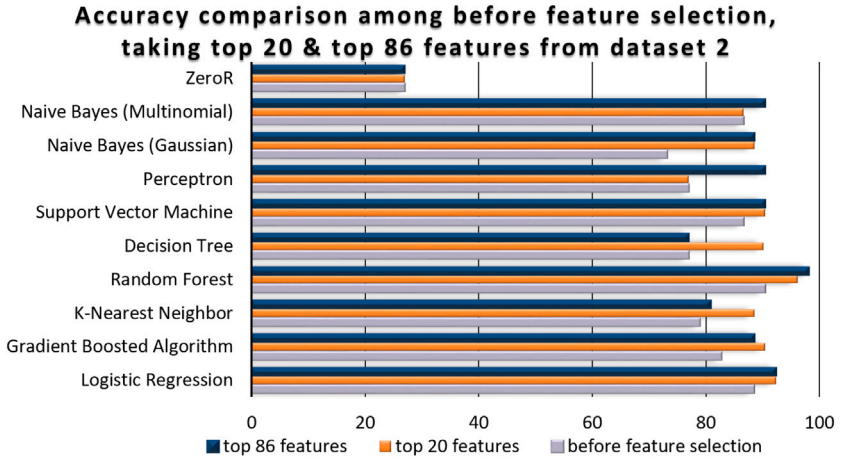
Table 9 compares the proposed depression detection system with similar works.

4.3. Cross validation on an independently collected dataset

One of the challenges with this model is that it is difficult to apply the model on other existing datasets. This is due to the fact that this model is trained on a dataset containing rich feature set while other existing dataset does not possess such extensive feature set. Therefore,



(a)



(b)

Fig. 12. Accuracies before and after feature selection (a) dataset 1 (b) dataset 2.

Table 8
Summary of the feature selection techniques for dataset 1 and dataset 2.

Machine learning algorithm	Dataset	Highest accuracy obtained	Highest accuracy (%)
ZeroR	1, 2	Same	26.92
Naive Bayes (Multinomial)	1	Considering top 20 features	67.31
Naive Bayes (Multinomial)	2	Considering top 86 features	90.38
Naive Bayes (Gaussian)	1	Considering top 20 features	61.54
Naive Bayes (Gaussian)	2	Considering top 86 features	88.46
Perceptron	1	Considering top 50 features	55.77
Perceptron	2	Considering top 86 features	90.38
Support Vector Machine	1	Same performance if top 20/top 50 features are considered	63.46
Support Vector Machine	2	Same performance if top 20/top 86 features are considered	90.38
Decision Tree	1	Considering top 50 features	67.31
Decision Tree	2	Considering top 20 features	90.08
Random Forest	1	Same performance if top 20/top 50 features are considered	65.38
Random Forest	2	Considering top 86 features	98.08
K-Nearest Neighbor	1	Considering top 20 features	63.46
K-Nearest Neighbor	1	Considering top 20 features	88.46
Gradient Boosting	1	Same performance if top 20/top 50 features are considered	67.31
Gradient Boosting	2	Considering top 20 features	90.38
Logistic Regression	1	Considering top 50 features	65.38
Logistic Regression	2	Same performance if top 20/top 86 features are considered	92.31

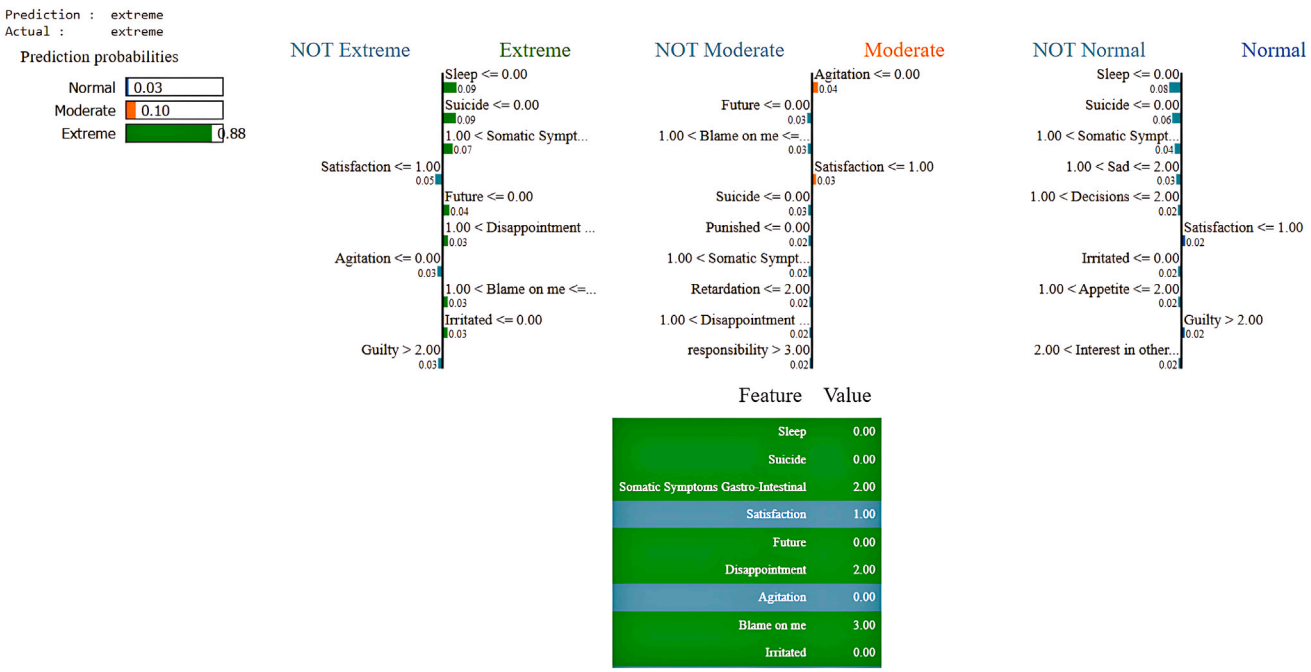


Fig. 13. Depression prediction interpretation of an extreme case of the LIME explainable AI.

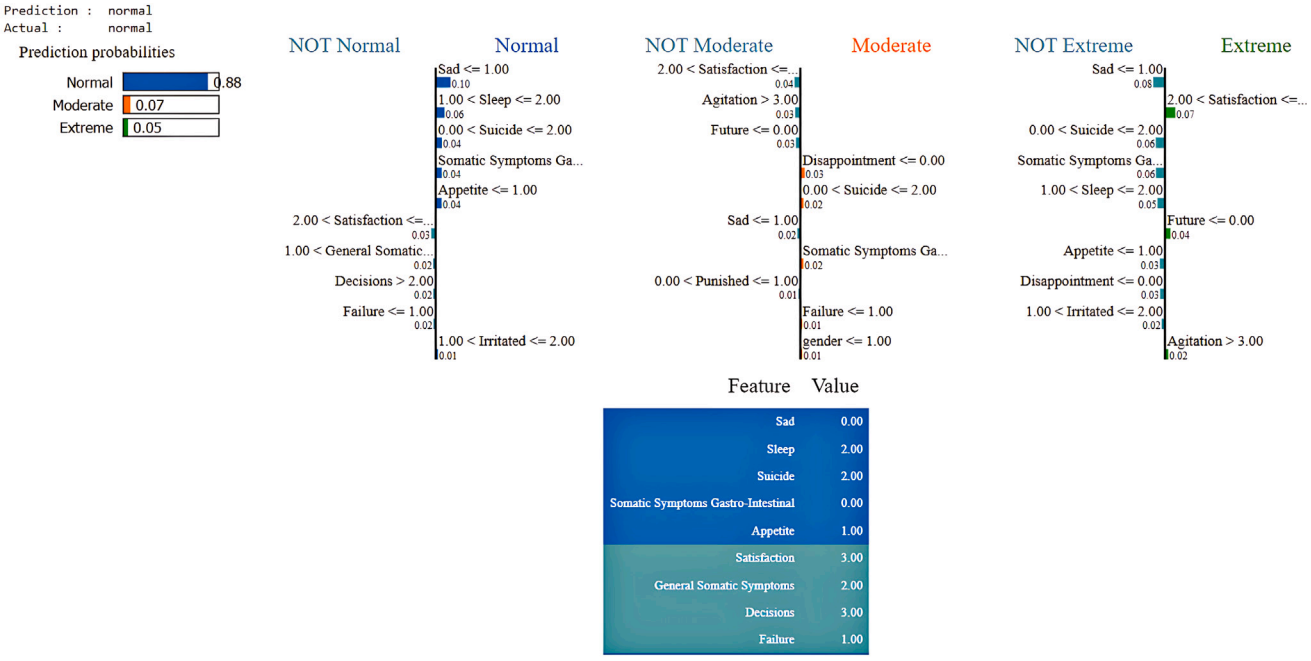


Fig. 14. Depression prediction interpretation of a normal instance of the LIME explainable AI.

in order to validate, we collected a new, independent dataset ($n = 74$) using the same features but from a different time period. The initial dataset was collected from 23 October, 2021 to 28 December, 2021 while this dataset was collected from 8 April 2023 to 18 April 2023. Four classifiers (the best performing ones) were applied on this dataset for cross validation and the results display similar performance on the cross validated dataset as that of the test dataset. Thus, the models have been found to be reliable in dataset collected from different time period. Please see Table 10 for detailed results.

5. Discussion

This study aims to identify the major reasons for depression among Bangladeshi university students. Our research found several factors exhibited by depressed university students. Most students are unaware of this mental disorder, and alarming 42.4% students consider committing suicide due to anxiety and stress. Experiencing emotional or sexual violence and financial hardship are discovered as important factors that contribute to depression. Besides, regular smoking, physical inactivity

Table 9
Comparison of the proposed depression prediction system with similar works.

Reference	Dataset	Number of participants	Used scale	Number of questions	Best model	Accuracy
[5]	Custom dataset of Bangladeshi students	210	Modified DASS-21	21	Binary logistic model	N/A
[11]	Undergraduate students from the University of Nice Sophia-Antipolis	4184	Electronic health records	59	XGBoost	AUC = 0.79
[12]	United States National Health and Nutrition Examination Survey (NHANES)	8628	PHQ-9	9	SVM	77.1%
[13]	Chinese Longitudinal Healthy Longevity Study (CLHLS)	1538	Information on health status and quality of life	8	Gradient Boosting	75.9%
[14]	Korean adults	623	National Institutes of Health Stroke Scale (NIHSS)	13	SVM	77.1%
[15]	Australian children	6310	Australian Child and Adolescent Survey	667	XGBoost	95%
[16]	Bangladeshi participants	6.4	Burns Depression Checklist (BDC)	55	AdaBoost	92.56%
[18]	Bangladeshi women	623	Lifestyle related questions	30	CNN	96.8%
This work	Custom dataset of Bangladeshi students	684	Combined scale	106	Random Forest	98.08%

Table 10
Model performance on test set and cross-validation dataset.

Classifier	Details on features/hyperparameters used	Test accuracy (%)	Accuracy on independent dataset (cross-val) (%)	Training time (ms)	Model size (KB)
Random Forest	Pearson Correlation (drop 20 features)	98.08	95.95	103.77	88.6
Gradient Boosting	RandomizedSearchCV (Hyperparameter Optimization)	94.23	93.24	210 755.27	88.6
Logistic Regression	GridSearchCV (Hyperparameter Optimization)	92.31	87.84	700.71	88.6
Logistic Regression	Recursive Feature Elimination (drop 20 features)	92.31	87.84	54.85	2.79

and frequent social media usage have been found to be common traits among depressed students.

Another objective of this research is to test if depression can be accurately predicted with the newly created depression assessment scale and using various machine learning and deep learning models. The new depression assessment questionnaires have been created employing the voting technique on eight well-known depression measuring scales. Various feature selection approaches and hyperparameter optimization techniques have been performed to enhance the performance of the prediction models. The accuracy is further increased by keeping the dominant features, i.e., removing irrelevant ones.

The model proposed here has been highly accurate and reliable compared to the notable works found in the literature, as demonstrated in Table 9. Empirical results reveal that machine learning and deep learning algorithms have been able to predict depression with high performance. The training duration and model size of the classifiers listed in Table 10 have been determined for the PC configuration shown below:

Processor: Intel (R) Core(TM) i7-6500U CPU @ 2.50 GHz 2.59 GHz, RAM: 8 GB, System Type: 64-bit operating system, x64-based processor.

Random forest has been found to produce the best result on the test set. Random forest uses ensemble approach that combines the strengths of various decision trees and hence it is likely to have provided a better result.

6. Limitation of the work

This article presents a novel approach of measuring depression as well as predicting depression using machine learning and deep learning models and was applied to Bangladeshi university going students. However, the work has experienced a number of challenges:

- The novel approach involved the use of a voting algorithm that was applied on eight well-known depression measuring scales. Consequently, this needed a much more extensive feature set compared to the previous work that exists in the literature.
- Due to extensive feature set, the questionnaire had a total of 106 questions which were time consuming for participants to fill up.
- Other datasets available in literature had much smaller set of features. As a consequence, it is not possible to apply our models on other existing datasets.
- This model was applied on university going students. Hence, it may not be reliable in inferring predictions on individuals with different profiles (such as children, elderly, etc...)

7. Conclusion and future scope

This research paper presents various machine learning and deep learning methods for assessing the level of depression among Bangladeshi university students. In addition, the study explores various personal and social factors that negatively impact young people's mental health. Moreover, this work has significant importance in studying

various factors of suicidal activities among young people that have been increasing recently, mostly due to depression. Our research created a new depression measuring scale with three distinct levels, i.e., normal, moderate and extreme, using the voting technique on the results of the eight recognized scales. A private dataset containing 684 participants has been collected following the consent of the participants. Next, nine feature selection methods were used to find relevant and most important features contributing to depression. In addition, 12 machine learning, ensemble and deep learning algorithms were applied to predict depression automatically. Random Forest, Gradient Boosting Algorithm and CNN were found to be the better models for evaluating depression. Finally, hyperparameter optimization and feature selection approaches were applied to enhance the efficiency of the prediction results. In the future, a comprehensive sample with more diverse cohorts can be combined with the existing dataset. Meta-heuristic optimization techniques can be applied to find the best features from patients' extensive biometric markers and characteristics. Possible extensions of this work are to utilize the prowess of more advanced artificial intelligence frameworks, e.g., adversarial and sequential learning, domain adaptation, etc.

CRediT authorship contribution statement

Rokeya Siddiqua: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Nusrat Islam:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft. **Jarba Farnaz Bolaka:** Conceptualization, Methodology, Investigation. **Riasat Khan:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Supervision. **Sifat Momen:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Dunn G, Sham P, Hand D. Statistics and the nature of depression. Psychol Med 1993;23:871–89.

[2] Organization WH. The World Health Report: Mental disorders affect one in four people. 2001.

[3] Organization WH. Depression. 2023, https://www.who.int/news-room/fact-sheets/detail/depression, [Online, accessed: 22 April 2023].

[4] Hossain MD, Ahmed HU, Chowdhury WA, Niessen LW, Alam DS. Mental disorders in Bangladesh: A systematic review. BMC Psychiatry 2014;14:1–8.

[5] Arusha A, Biswas R. Prevalence of stress, anxiety and depression due to examination in Bangladeshi youths: A pilot study. Child Youth Serv Rev 2020;116:1–6.

[6] Chang J, Yuan Y, Wang D. Mental health status and its influencing factors among college students during the epidemic of COVID-19. J South Med Univ 2020;40:171–6.

[7] Choudhury AA, Khan MRH, Nahim NZ, Tulon SR, Islam S, Chakrabarty A. Predicting depression in Bangladeshi undergraduates using machine learning. In: IEEE region 10 symposium. 2019, p. 789–94.

[8] Hoque R. Major mental health problems of undergraduate students in a private university of Dhaka, Bangladesh. Eur Psychiatry 2015;30:1880.

[9] Sultana J, Elhum Uddin Quadery S, Amik FR, Basak T, Momen S. A data-driven approach to understanding the impact of Covid-19 on dietary habits amongst Bangladeshi students. J Posit Sch Psychol 2022;6:11691–7.

[10] Conceição V, Rothes I, Gusmão R. The association between stigmatizing attitudes towards depression and help seeking attitudes in college students. PLOS ONE 2022;17:1–14.

[11] Nemesure M, Heinz M, Huang R, Jacobson N. Predictive modeling of psychiatric illness using electronic health records and a novel machine learning approach with artificial intelligence. Sci Rep 2020;11:1–9.

[12] Lee C, Kim H. Machine learning-based predictive modeling of depression in hypertensive populations. PLOS ONE 2022;17:1–17.

[13] Su D, Zhang X, He K, Chen Y. Use of machine learning approach to predict depression in the elderly in China: A longitudinal study. J Affect Disord 2021;282:289–98.

[14] Ryu YH, et al. Prediction of poststroke depression based on the outcomes of machine learning algorithms. J Clin Med 2022;11.

[15] Haque UM, Kabir E, Khanam R. Detection of child depression using machine learning methods. PLOS ONE 2021;16:1–13.

[16] Zulfiker MS, Kabir N, Biswas AA, Nazneen T, Uddin MS. An in-depth analysis of machine learning approaches to predict depression. Curr Res Behav Sci 2021;2:1–12.

[17] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artificial Intelligence Res 2002;16:321–57.

[18] Ahmed A, Sultana R, Ullas MTR, Begom M, Rahi MMI, Alam MA. A machine learning approach to detect depression and anxiety using supervised learning. In: Asia-Pacific conference on computer science and data engineering. 2020, p. 1–6.

[19] Moon NN, Mariam A, Sharmin S, Islam MM, Nur FN, Debnath N. Machine learning approach to predict the depression in job sectors in Bangladesh. Curr Res Behav Sci 2021;2:1–10.

[20] Cassidy S, Bradley L, Bowen E, Wigham S, Rodgers J. Measurement properties of tools used to assess depression in adults with and without autism spectrum conditions: A systematic review. Autism Res 2018;11:738–54.

[21] Ustun G. Determining depression and related factors in a society affected by COVID-19 pandemic. Int J Soc Psychiatry 2021;67:54–63.

[22] Saha A, Dutta A, Sifat RI. The mental impact of digital divide due to COVID-19 pandemic induced emergency online learning at undergraduate level: Evidence from undergraduate students from Dhaka City. J Affect Disord 2021;294:170–9.

[23] Bathla M, Singh M, Kulhara P, Chandna S, Aneja J. Evaluation of anxiety, depression and suicidal intent in undergraduate dental students: A cross-sectional study. Contemp Clin Dent 2015;6:215–22.

[24] Ibrahim AK, Kelly SJ, Glazebrook C. Reliability of a shortened version of the Zagazig Depression Scale and prevalence of depression in an Egyptian university student sample. Compr Psychiatry 2012;53:638–47.

[25] Guo T, Guo Z, Zhang W, Ma W, Yang X, Yang X, Hwang J, He X, Chen X, Ya T. Electroacupuncture and cognitive behavioural therapy for sub-syndromal depression among undergraduates: A controlled clinical trial. Acupunct Med 2016;34:356–63.

[26] Ozawa C, et al. Resilience and spirituality in patients with depression and their family members: A cross-sectional study. Compr Psychiatry 2017;77:53–9.

[27] McIntyre RS, et al. The prevalence and illness characteristics of DSM-5-defined "mixed feature specifier" in adults with major depressive disorder and bipolar disorder: results from the International Mood Disorders Collaborative Project. J Affect Disord 2015;172:259–64.

[28] Burchert S, Kerber A, Zimmermann J, Knaevelsrud C. Screening accuracy of a 14-day smartphone ambulatory assessment of depression symptoms and mood dynamics in a general population sample: Comparison with the PHQ-9 depression screening. PLoS One 2021;16:1–25.

[29] Wang K, et al. Mapping of the acromegaly quality of life questionnaire to ED-5D-5L index score among patients with acromegaly. Eur J Health Econ 2021;1–11.

[30] Mergen H, et al. Comparative validity and reliability study of the QIDS-SR16 in Turkish and American college student samples. Klin Psikofarmakol Bülteni-Bull Clin Psychopharmacol 2011;21:289–301.

[31] Lu S, Hu S, Guan Y, Xiao J, Cai D, Gao Z, Sang Z, Wei J, Zhang X, Margraf J. Measurement invariance of the Depression Anxiety Stress Scales-21 across gender in a sample of Chinese university students. Front Psychol 2018;9:2064.

[32] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: International conference on knowledge discovery and data mining. 2016.