1236: EXPLAINABLE ARTIFICIAL INTELLIGENCE SOLUTIONS FOR IN-THE-WILD HUMAN BEHAVIOR ANALYSIS

Enhancing explainability in predicting mental health disorders using human–machine interaction

Inderpreet Kaur[1] · Kamini[1,3] · Jaskirat Kaur[1,4] · Gagandeep[1] · Simar Preet Singh[2] · Umesh Gupta[2]

Abstract

Mental health measures an individual's emotional, psychological, and social well-being. It influences how a person thinks, feels, and responds to events. Mental illness has wreaked havoc on society in today's globe and has come to the forefront as a serious concern. People with mental disorders, including bipolar disorder, schizoaffective disorder, sadness, anxiety, and others, rarely recognize their condition as the world's most serious problem. In mental illness, there are a variety of emotional and physical symptoms. Anxiety attacks, sweating, palpitations, grief, worry, overthinking, delusions, and illusions are all symptoms of mental illness, and each symptom indicates the kind of mental disorder. Our study outlined the standardized approach for diagnostic depression, including data extraction, pre-processing, ML classifier training, identification classification, and performance assessment that enhances human–machine interaction. This study utilized five machine learning methods: k-nearest neighbor, linear regression, gaussian classifier, random forest, decision tree, and logistic regression. The accuracy, precision, recall, and F1-score metrics are used to evaluate the efficacy of machine learning models. The algorithms are categorised according to their accuracy, and explainability shows that the Gaussian classifier (Minmax scaler), which reaches 91 per cent accuracy, is the most accurate. Furthermore, given that the characteristics are predicated on potential indications of depression, the approach is capable of producing substantial justifications for the determination via machine learning models employing the SHapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) algorithms of explainable Artificial Intelligence (XAI). Thus, the approach to predicting depression can aid in the advancement of intelligent chatbots and other technologies that improve mental health treatment.

Keywords Mental health · Machine learning · Predictive models · Gaussian classifier · SHAP · LIME

Extended author information available on the last page of the article

1 Introduction

Mental health and well-being are essential aspects of life quality; they are necessary components of social cohesion, productivity, and the maintenance of stability and peace in the living environment [1]. Mental illness has caused havoc on society in today's globe and has come to the forefront as a serious concern. People who suffer severe depression, anxiety, panic disorder, schizophrenia, bipolar disorder, and other mental problems are mostly oblivious that mental illnesses constitute the world's most serious issue [2]. Because it produces disturbance in a person's behavior, thoughts, and emotions, it has a detrimental impact on their life. Mental illness leads to various changes and bad outcomes in a person's life, such as losing trust in himself and his family. Mental illness has social ramifications and a monetary cost on a worldwide scale [3]. There are no biological tests available or done to assess mental disorders. The expert's opinion is employed to determine the diagnosis based on the various symptoms. There may be an obvious link and effect based on the person's physiological symptoms. Mental health and a more positive mindset are essential for one's entire well-being. Each year, eight lakhs' people attempt suicide, by estimates, with a much higher number of persons committing suicide. The majority of suicides are caused by prevalent mental diseases [4]. Suicide is a top cause of mortality worldwide and a leading cause of adolescent death. Schizophrenia, bipolar disorder, and chronic disorders suddenly manifest as serious mental health issues. Such conditions may be avoided or treated more successfully. More treatment and care may be offered if abnormal mental states are identified early in the disease [5]. Furthermore, diagnosis-based methods have the unintended effect of inhibiting the participation of those who are ill. As a consequence, psychological issues are usually overlooked or ignored. Anxiety and depression issues have a variety of health and well-being consequences. Ischemic heart disease, hypertension, diabetes, accidental accidents, and purposeful accidents are risk factors for anxiety and depression. In addition, insomnia, palpitations, tremors, severe weight loss or increase, diarrhea, and vomiting are a few symptoms. Explainability draws attention to the close connection between suicidal ideas and sadness, where hopelessness may tragically lead to suicide. Unfortunately, those who suffer from depression and anxiety often experience discrimination from both society and their family, which emphasises the need for explainability in combating such prejudices. As a result, people could have difficulties in both academic and professional contexts; explainability sheds light on these difficulties. Their problems are made worse by economic stress, which is sometimes difficult to detect, resulting in poverty and deteriorating health. The deterioration of economic and social changes further lowers their standard of living. Financial stress disproportionately affects low- and middle-income families [6].

In recent years, data collection in mental health therapy has increased substantially. The hearing may be impaired as a consequence of the depressed mood. People behave in strange ways in general. Mental disease is on the increase nowadays. Everyone else has a mental disease of some kind. However, only 35 to 50 percent of people in high-income countries who need primary care actually get it, highlighting the importance of explainability in determining the best course of action. This imbalance differs since diagnostic treatment is restricted, and the disease is not detected first and foremost [7]. Social media, smartphones, neuro-imaging, and wearables enable mental health researchers to gather data swiftly. Machine learning uses probabilistic and statistical methodologies to create self-learning machines. AI (artificial intelligence), natural language processing, speech recognition, and computer vision have all benefited from machine learning. In addition,

machine learning has helped significant progress in fields like bioinformatics by enabling the rapid and scalable examination of complex data. Similar analytic approaches are being used to study mental health data to improve patient outcomes and improve knowledge of psychiatric illnesses and their treatment [8, 36]. We employed machine learning and data scaling methods to identify mental disabilities and fill these gaps. Consequently, the following provides a summary of our technological contribution:

The machine learning-based technical article made the following significant contributions:

- The main goal of our research project is to provide a better method for mental illness prediction.
- We used six machine learning algorithms, including data scaling methods like min–max and standard scaling, to standardise the mental illness dataset.
- We thoroughly examined the algorithms' performance with and without these scaling techniques. After analysing the results of our investigation, we concluded that the Gaussian classifier was the most effective method.
- The applied methodologies used quality criteria such as F1 score, accuracy, precision, and recall to predict mental health disorders.

The remainder of the paper is as follows: The evaluation of the background investigation has been summarised in Section 2, which contains a comprehensive analysis of the previous results. We used a variety of measures to evaluate the model's performance, including each classifier's F1 score, accuracy, precision, and recall. The details of this assessment are provided in Section 3. Section 4 then explores the preparation steps, including data manipulation and visualisation. Afterwards, the Section 5 explores the machine learning classifiers used most often. The Section 6 provides the results and a comparative analysis of these classifiers, along with recommended reading lists. Section 7 concludes by summarising the results of our investigation and offering suggestions for further research efforts.

2 Background study

Sau et al. [9] focused on using machine learning techniques in the realm of automated mental health disease screening. Using this technology, an automated computer-based method may replace time-consuming anxiety and depressive episodes with acceptable accuracy. With 82.6 percent accuracy and 84.1 percent precision, Catboost proved to be the best choice for this task. Braithwaite et al. [10] used machine learning techniques using Twitter data to verify suicidality measurements in the US population. The results reveal that machine learning algorithms can effectively detect the clinically severe suicide rate in 92% of instances. Srividya et al. [11] developed a model for measuring mental health using machine learning techniques. Clustering was employed before constructing models. MOS was used to validate and train a classifier's class labels. The use of categorization models improved psychological health prediction accuracy to 90%. Watts et al. [12] employed machine learning, and various historical, sociodemographic, and clinical factors were assessed using data-driven methods. Sexual crimes may be predicted from nonviolent and violent crimes using 36 factors. A binary classifier has an 83.26% sensitivity and a 76.42% specificity for predicting sexual and violent offenses. Several algorithms were developed using the data of 970 participants, according to Hornstein et al. [13], to forecast

a substantial decrease in depressive and anxious symptoms utilizing clinical and sociode-mographic characteristics. First, it was decided to forecast the results of 279 newcomers using an RF (random forest) classifier since it outperformed cross-validation the best. Jain et al. [14] state that the suggested model has used eight widely used machine learning (ML) calculation techniques to build up the expectations models using a large dataset, resulting in exact This study endeavour has made an effort to obtain a precise and exact image by using a variety of methodologies and models. Following different strategies helps to clarify information, do tasks more effectively, and decrease the frequency of suicide cases. The Support Vector Machine was used, and the ultimate result was 87.38 percent (SVM).

Twelve percent of new mothers suffer from PPD (postpartum depression), a dangerous medical condition, as per Andersson et al. [15]. Preventive therapies are most beneficial for high-risk women, but it may be challenging to identify them. The randomized tree approach performed well, providing the highest accuracy and a balanced sensitivity and specificity. In order to predict psychological discomfort from just ecological parameters, Sutter et al. [16] built a machine learning model using a variety of methodologies. On a sample dataset, eight different classification approaches were used and achieved accuracy of 0.811. The early findings imply that an accurate and trustworthy model is feasible with further advancements in implementation and analysis. With the suggested basic paradigm, this research may help create a proactive response to the world's mental health epidemic.

Rahman et al. [18] state that to find meaningful features, six distinct feature selection algorithms were used to sort through the 34 characteristics retrieved from each signal. According to empirical research, a neural net (NN) can distinguish between the three musical genres with 99.2% accuracy using a collection of characteristics taken from physiological data. The research also finds a few helpful factors to boost classification model accuracy. Anxiety and depression are common among the elderly, as per Sau et al. [18]. This article uses machine learning to predict anxiety and sadness in elderly individuals. Ten classifiers were evaluated using ten-fold cross-validation on 510 elderly individuals and achieved the highest accuracy of 91%. Sano et al. [19] developed wearable sensors, mobile phones, and data integrity techniques for objective behavioral and physiological assessments Classification accuracy for stress was 73.5% (139/189), and for mental health, it was 79% (37/47) using modifiable behavioral factors. Table 1 describes the comparative analysis of the past study.

3 Proposed design

This section discusses the potential approaches and datasets for early mental disease prediction. Python and Jupyter notebook's integrated design environment (IDE) was used for all the implementation work. As shown in Fig. 1, the suggested framework aims to increase the precision of early mental disease prediction. The information was accessible on the Kaggle website. The study's data source is a single CSV file with a string data type containing different diseases. The data was first translated into a Boolean form from yes or no questions depending on the signs of different diseases. The missing values were confirmed once the dataset was imported into Python as a Data frame. Pre-processing included converting the string datatype into the 0 and 1-based Boolean type. After completing the cleaning procedure, we applied a feature selection technique to limit the features and selected just the finest.

The following machine learning techniques are used to categorize the illnesses into six groups once the dataset has been imported. The pre-processing & feature selection stages

Table 1 Comparative Analysis

Reference	Datasets	Techniques Used	Limitations	Findings
[9]	Questionnaire	Logistic regression (LR) naïve bayes (NB), support vector machine (SVM), random forest (RF), catboost (Cat)	Scope is confined to the employed population	The most excellent accuracy and precision obtained by Catboost are 82.6 percent and 84.1 percent, respectively
[10]	Data taken from social media	Decision tree (DT)	Limited significant clinical suicidality was investigated	DT model achieved accuracy of 91.9%
[11]	Questionnaire	Support vector machines, decision trees, naïve bayes (NB), K-nearest neighbor, logistic regression (LR) and ensemble model	More mental disease categories in the target group means greater class labels, with more overlap	Ensemble learning model outperformed with accuracy of 90%
[12]	Data obtained from ORB (2014–2015)	Support vector machine, random forest, ElasticNet	Individual crime categories were differentiated using binary classifiers only	RF had outperformed with an accuracy of 80.31%
[13]	Data collected from from Meru Health program	Random forest	Limited sample size	The accuracy of the random forest classifier was 0.71, with an AUC of 0.60
[14]	Data acquired from survey	Random forest (RF), support vector machine (SVM), naïve bayes (NB), decision tree (DT), logistic regression (LR), XGBoost (XGB), gradient boosting classifier (GBC) and artificial neural network (ANN)	This method requires additional improvement in categorization accuracy	SVM achieved the highest accuracy of 87.38%
[15]	Data collected from project	Ridge regression (RR), LASSO regression, gradient boosting machines (GBM), distributed radom forests (DRF), naïve bayes and stacked ensembles models	sources of bias in selection, Class disparity	Extreme randomized forests outperformed with accuracy 73%, AUC of 81%
[17]	Data collected from 20 participants	Neural network	Limited dataset	Model achieved accuracy of 98.5%

Table 1 (continued)

| Reference	Datasets	Techniques Used	Limitations	Findings
[18]	Data taken from patients	J48, bayesian network (BN), logistic, multiple layer perceptron (MLP), Naïve bayes (NB), random forest (RF), random tree (RT), sequential minimal optimisation (SMO), BN, logistic, MLP	Study did not identify elderly people who had underlying mental health issues	Random forest outperformed with 91% accuracy
[19]	Data obtained from surveys	Support vector machine, least absolute shrinkage and selection operator (LASSO)	Distinguishing between two categories will not really indicate it's relevant or causes the behaviour, 64% of research participants were men	Model has achieved accuracy of 87%
[20]	Data colllected from citizens of bangladesh	K nearest neighbor, AdaBoost, gradient boosting (GB), extreme gradient boosting (XGBoost), bagging	No biological marker for depression was provided	AdaBoost accuracy-92.56%
[21]	Data obtained from UCLA-xDR	CART modelling	This approach needs significant classification precision enhancement	Model achieved accuracy of 0.84, AUC-0.73, sensitivity-73.4
[22]	Data acquired from medical centres	Support vector machine (SVM)	Proposed technique handles training or misclassification mistakes at a low rate	SVM achieved accuracy of 0.825, sensitivity-90%, and specificity -74%
[23]	Collected dataset from BAVED	Convolutional neural network (CNN), Support vector machine (SVM)	Classification method needs to be improved	Hybrid model (CNN + SVM) obtained accuracy of 81.60%, sensitivity-98.5%, specificity-81.6%
[24]	Taken data from published sources	Generalized linear model, kNN, SVM, RF, RT	Limited dataset, overfitting problem	SVM outperformed with 83% accuracy
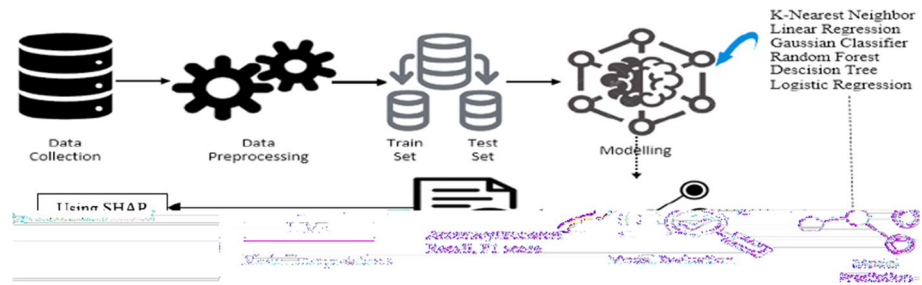
Fig. 1 Proposed framework

have been finished: The gaussian classifier, the decision tree (DT), the linear regression, the logistic regression (LR), the random forest (RF), the K-nearest neighbor (KNN), and (LR). SHAP and LIME explainers have been incorporated into this paper. The implementation of these models represents an advanced approach to augment the visibility of machine learning (ML) models. They offer insights from both a regional and international standpoint, detailing the individual impact of each factor on the ultimate probability linked to the possible emergence of the pathology. A detailed elucidation of the various machine learning (ML) techniques and SHAP and LIME explainers used in the investigation can be found in Sect. 4. Moreover, by using a range of assessment criteria, the algorithms are examined and contrasted in terms of their capacity to forecast future events (as delineated in the findings section), highlighting the significance of explainability in the assessment process.

3.1 Dataset description

It is crucial to building a trustworthy model and fully comprehend every aspect of it before employing the model. It could be immoral to obtain the necessary data from a few sources, even if it may be available from many. As a result, information from reliable sources could not be trustworthy sometimes, which calls for much ethical work. In order to go forward with this study, we need data that includes different signs of mental illness. Research indicates that a person's symptoms are considered when diagnosing mental diseases. The study utilised the "mental health prediction" dataset from Kaggle (https://www.kaggle.com/code/kairosart/machine-learning-for-mental-health-1). First, we look at the dataset, which has 334 rows and 31 columns. The collection includes a variety of mental illnesses and their symptoms, and each illness has a wide range of symptoms.

There are 31 columns in the dataset. Some attributes include feeling anxious, depressed, anxious, panic attacks, mood swings, difficulty concentrating, difficulty sleeping, hopelessness, anger, tiredness, trouble concentrating, compulsive behavior, etc. These signs and symptoms depend on the kind of mental disease. This dataset's primary goal is to identify connections between people's mental illnesses and their occupations and levels of education. For mental illness, many criteria have been found. The dataset, along with the explanation, is shown in Fig. 2.

3.2 Exploratory data analysis

Data visualization is the graphic depiction of data used to comprehend its relationships and patterns. Building machine learning models is significantly more accessible because of EDA (Exploratory Data Analysis), which enables better knowledge of the data. Plots, a data visualization tool based on Matplotlib, has been used to examine

```
In [17]:  df.info()

          <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 333 entries, 1 to 333
          Data columns (total 31 columns):
           #   Column                                                Non-Null Count  Dtype
          ---  ------                                                --------------  -----
           0   I am currently employed at least part-time            333 non-null    in:64
           1   MentallyIll                                           333 non-null    in:64
           2   Education                                             333 non-null    object
           3   I have my own computer separate from a smart phone    333 non-null    in:64
           4   I have been hospitalized before for my mental illness 333 non-null    in:64
           5   How many days were you hospitalized for your mental illness  333 non-null    float64
           6   I am legally disabled                                 333 non-null    in:64
           7   I have my regular access to the internet              333 non-null    in:64
           8   I live with my parents                                333 non-null    in:64
           9   I have a gap in my resume                             333 non-null    in:64
           10  Total length of any gaps in my resume in months.     333 non-null    in:64
           11  Annual income (including any social welfare programs) in USD  333 non-null    in:64
           12  Unemployed                                            333 non-null    in:64
           13  I read outside of work and school                     333 non-null    in:64
           14  Annual income from social welfare programs            333 non-null    in:64
           15  I receive food stamps                                 333 non-null    in:64
           16  I am on section 8 housing                             333 non-null    in:64
           17  How many times were you hospitalized for your mental illness  333 non-null    in:64
           18  Lack of concentration                                 333 non-null    float64
           19  Anxiety                                               333 non-null    in:64
           20  Depression                                            333 non-null    in:64
           21  Obsessive thinking                                    333 non-null    float64
           22  Mood swings                                           333 non-null    float64
           23  Panic attacks                                         333 non-null    float64
           24  Compulsive behavior                                   333 non-null    float64
           25  Tiredness                                             333 non-null    float64
           26  Age                                                   333 non-null    object
           27  Gender                                                333 non-null    object
           28  Household Income                                      333 non-null    object
           29  Region                                                333 non-null    object
           30  Device Type                                           333 non-null    object
          dtypes: float64(7), int64(18), object(6)
          memory usage: 80.8+ KB
```

Fig. 2 Detailed description of dataset

data visualization approaches in this research [40]. The EDA facilitates the detection of hidden characteristics in data collection. It includes a sophisticated user interface that enables you to design eye-catching and educational statistics visuals.

Several libraries were loaded, namely NumPy, Matplotlib, Seaborn, and Pandas, to perform the EDA. Pre-processing included converting the string datatype into the 0 and 1-based Boolean type. We can get the basic information about the Dataframe using Dataframe.info (). dataframe is a mix of categorical and numerical data. We also have the missing values in some columns and we can get some more insight about the data using Dataframe.describe (). In EDA, we have digged into data and find information about the no of people who have survived with mental illness. So, 122 out of 611 survived with mental disorder and 250 have not survived with mental illness as plotted in Fig. 3(a). In Fig. 3(b), We utilized many characteristics, such as high school, undergraduate, completed masters and completed PhD, to depict the data for various education levels along with their count.

3.3 Data transformation

It is essential to deal with the number of missing values in the data. So, in this step, we are to find the missing data because data is usually taken from multiple sources, which
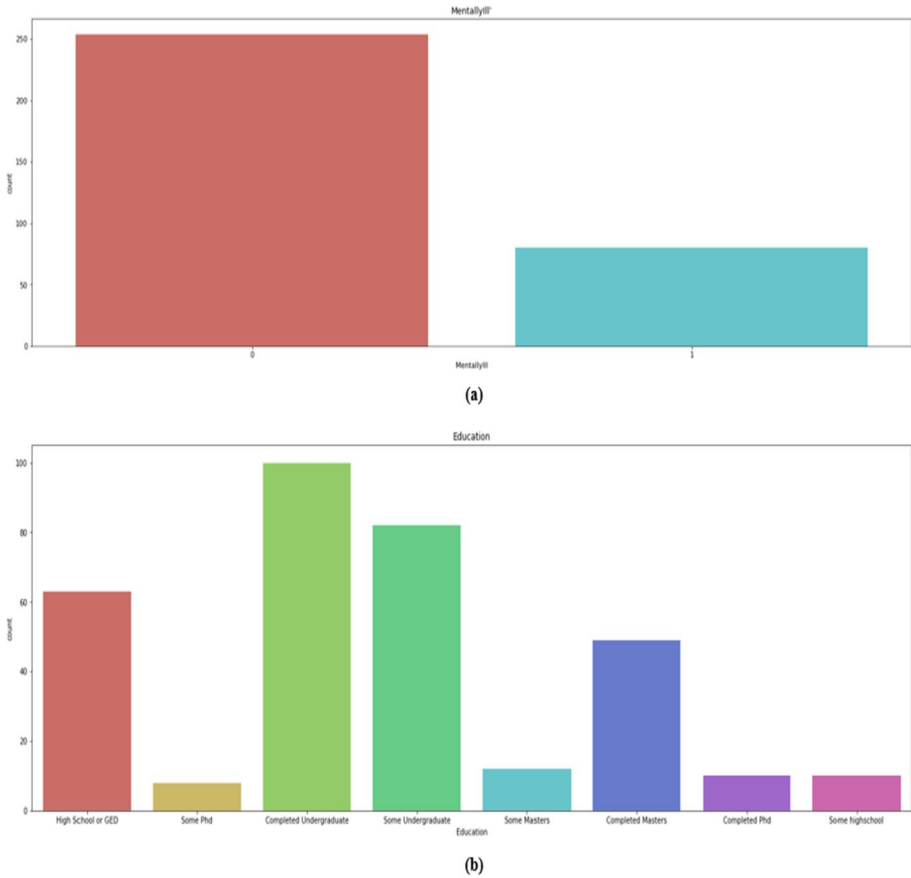
Fig. 3 (**a**) Plotting of people having and not having mental illness (**b**) Plotting of people having different types of mental illness

are generally not reliable. To tackle the missing values, we first consider them by drop column methods and estimate the missing values with the mean, median, or mode of the respective features. Datasets can have inconsistent values. To deal with conflicting values, it is necessary to perform data assessment like knowing the data type of a particular column and checking its correspondence rows values. We have replaced the column value responses to the 0 and 1. where "0" means No and "1" means yes. We have applied the numeric function to convert the income values into numeric. First, 1 and 0 have been used in place of the yes and no. Additionally, we must substitute Nan with 0 in the next step. For this, we utilized the fillna() method. Second, we have substituted missing data with imputation. Imputation seeks to replace missing data values with another value, often the mean or median of the given column, instead of discarding these values. The database has some missing values, which are shown as NaN. Estimating or impugning missing values using the available data is a more effective method. We used the sklearn, SimpleImputer algorithm to impute missing data. A Simple Imputer object must first be fitted to the training set in order to learn only a few parameters, such as the mean, median, and mode; otherwise, when applied to the whole dataset,
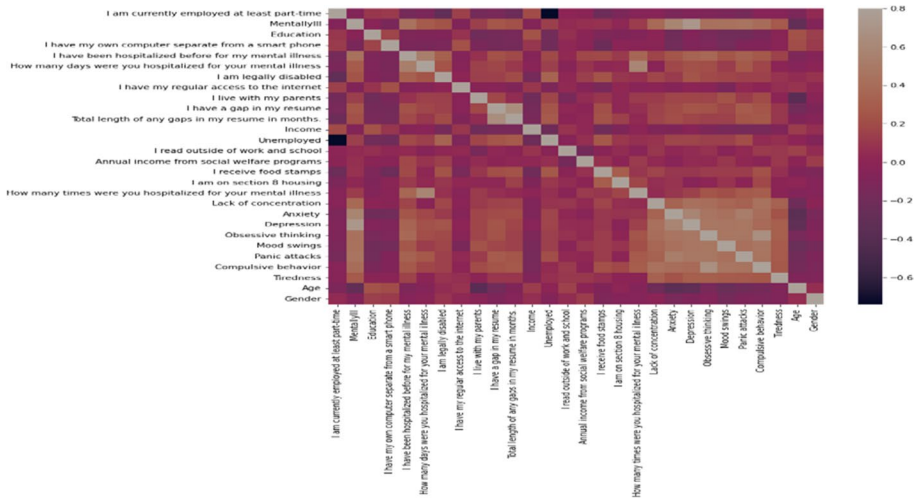
Fig. 4 Correlation Matrix

it would provide erroneous results. We used the parameters discovered from training set to impute missing information on the testing set. To use a categorical variable in a machine learning model, one must "encode" it into a numerical value. Next, we encoded target labels ranging from 0 to n class-1 using Label Encoder on our study dataset [41]. This transfer should be used to encode the target values.

3.4 Covariance matrix

Correlation between two variables shows how much those two variables contain the same information. Correlation is significant for learning more about data. The variances of the variables will be along the main diagonal of the variance–covariance matrix, and the covariances between any two variables will be in the remaining matrix cells. The means of the variables are also represented in the mean vector. [25, 26]. In our study, we have used a correlation between mental illness with job description and education. In the Fig. 4, Strong negative correlation between obvious features of unemployed and currently employed at least part time. Strong positive correlation between depression and mentally ill, attacks and anxiety.

3.5 Feature selection

The technique for minimizing the components to support an advanced model is feature selection. Feature selection techniques aim to enhance model accuracy and minimize computation cost by eliminating superfluous features and making models and procedures easier to understand and more suited to a dataset [26]. It also helps to minimize training length, data complexity, and model overfitting. The selection of features is a preprocessing step used prior to establishing a classification model that addresses the curse of dimensionality, which adversely impacts the algorithm. For feature classification we have used parameters as shown in Fig. 5.
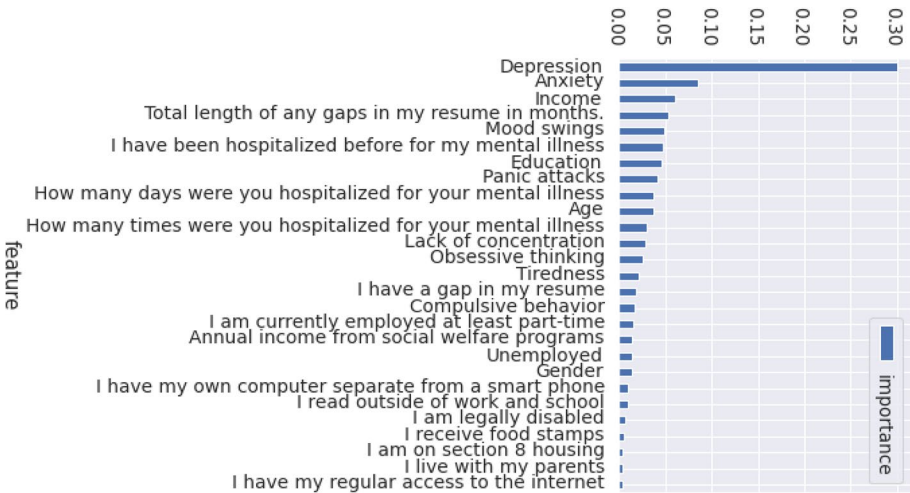
Fig. 5 Features used in study

3.6 Scaling methods (standard scaler and minMax scaler)

Data transformation consists of all operations we need to change the variables or create new ones. Dataset has some entries with different scales that can cause issues while training the model. We employed both min–max & standard scalers in this research study. Observed values of a variable are rescaled such that their new distribution has a mean of zero and a standard deviation of one using a standard scaler as shown in Eq. (1).

$$y = \frac{x - x_{mean}}{x_{std}} \tag{1}$$

where y is the scaled data and x is to be scaled data. Here, x_{mean} is the mean of the training samples and x_{std} is the standard deviation of the training samples [27]. To contrast, Min–max scaling uses the following formula to scale each characteristic independently as shown in Eq. (2).

$$y_i' = 2 \frac{y_i - y_{min}}{y_{max} - y_{min}} - 1 \tag{2}$$

where y_i is the log(y) response's initial value and y' is its scaled value at depth i. y_{min} and y_{max} are the log response's minimum and maximum values, respectively [28].

4 Enhancing mental health prediction using explainable machine learning

Incorporating explainability within Predicting Mental Health Disorders via the Human–Machine Interaction framework is a crucial undertaking. This research study explores machine learning (ML) approaches, such as k-nearest neighbor, linear regression,

Gaussian classifier, random forest, decision tree, and logistic regression. By strategically using these machine learning approaches, the project aims to provide precise forecasts and prioritize elucidating the underlying mechanisms of the models used for prediction. Each technique brings forth its distinct aspect: the k-nearest neighbor method enables the examination of data through proximity analysis, linear regression establishes patterns for prediction, the Gaussian classifier introduces a probabilistic understanding, decision trees and random forest algorithms dissect complex decision pathways, and logistic regression provides a probabilistic basis for classification. Through the strategic use of several methodologies, this study aims to improve the accuracy of predictions related to mental health disorders. Additionally, it seeks to understand better the underlying processes that contribute to these predictions. The objective of this work is to provide a better experience and productive collaboration between humans and artificial intelligence in the realm of mental health condition prediction.

In order to get insight into disease outbreaks, several machine learning approaches may help uncover patterns of abnormal behavior within data. This section describes well-known machine learning methods for predicting mental illness.

- **Logistic regression**

The sigmoid function graph serves as a visual representation of logistic regression. It was given the logistic function's name. A point may represent any real number on S-shaped curve, which is always somewhere between zero and one [29]. A logistic regression equation as shown in Eq. (3) is used in statistical software to comprehend the link between a dependent variable & one or more independent variables by measuring probabilities.

$$p_i = f\left(\frac{y}{x}\right) = \frac{1}{1 + exp\{z\}} \tag{3}$$

Logistic regression may be expanded to include several independent variables, which can be either continuous or categorical. The predicted probability of an event occurring is denoted by p_i in the preceding calculation. In such case, the multiple logistic regression approach would be:

$$p_i = \frac{1}{1 + exp\{-z_i\}} = \frac{exp\{z_i\}}{1 + exp\{z_i\}} \tag{4}$$

where p_i is the response variable and exp $\{-z_i\}$ are regression coefficients that express the proportional change in y for a unit change in z_i.

- **K-Nearest neighbor**

One non-parametric supervised learning method used for classification and regression is K-NN. The value assigned to a data instance in the primary K-closest neighbors classification is decided by the simple majority vote of the nearest neighbors [30]. The classification uses uniform weights. Alternately, the weights might be computed using a user-defined distance function. Analyzing neighborhood components is similar to studying the Mahalanobis distance measure.

$$L(xi - xj)||2 = (xi - xj)TM(xi - xj) \tag{5}$$

where $M = L^t L$ is a symmetric positive semi definite matrix of size (n_features, n_features).

- **Decision Tree**

Decision tree classifier is a supervised machine learning algorithms which uses a set of rules to make decision like human make decisions. Here, we can used classification and regression model to make decisions. The main idea behind using of decision trees classifiers is to use the dataset features to provide answer of yes/no questions and then split the dataset into data points of each class [31]. If a target is a classification outcome taking on values 0,1,…,K-1, for node m,

$$p_{mk} = 1/N_m \sum_{y \epsilon Qm} 1(y = k) \tag{6}$$

Let be the proportion of class k observations in node m. If m is a terminal node, predict_proba for this region is set to p_{mk}.

- **Random Forest**

A meta estimator called a random forest fits several decision tree classifiers to various dataset subsamples. As its name suggests, a random forest is an ensemble of several independent decision trees. Each distinct tree generates a class prediction, and our model's prediction is determined by the class that receives the most votes. This strategy minimizes the estimator's variance and avoids overfitting by merging many trees with a little more significant bias [32].

- **Gaussian Classifiers**

The Gaussian Processes Classifier is a ML classification approach and an extension of the Gaussian probability distribution that may serve as the basis for advanced non-parametric machine learning classification classification and regression algorithms. To a large extent, Gaussian classifiers depend on the Mahalanobis distance, which is the basis of their distance technique [33]. Due to the equal summation of squared distances across all features, this distance measure decreases in noise despite its widespread use.

$$p(x \lor y = c) = N \tag{7}$$

Here, Mean and covariance matrix for a Gaussian distribution with N independent variables are denoted by $p(x \lor y = c)$

- **Linear Regression**

The linear regression method determines how well the value of the independent variable influences the significance of the dependent variable by expressing a linear connection between a dependent variable (Y) & one or more independent (X) variables [34]. First, the coefficient represented by the Greek letter β(Beta), also known as a linear equation scaling factor, is applied to the input numbers.

$$Y = \beta 0 + \beta 1 * X \tag{8}$$

Multiple Linear Regression, with βo being the intercept β1,β2,β3,β4…,βn, and X1, X2, X3, X4, Xn being the independent variables' coefficients or slopes, and Y being the dependent variable

$$Y = \beta 0 + \beta 1 * X1 + \beta 2 * X2 + \beta 3 * X3 + \dots \dots + \beta n * Xn \tag{9}$$

Linear Regression models utilize to determine the straight line that best fits the data and the corresponding intercept and coefficient values that will result in the slightest possible error.

5 Basic concepts of SHAP and LIME

Providing light on machine learning's underneath is crucial for building confidence among healthcare professionals and guaranteeing openness in the machine learning-driven decision-making process. Thus, in order to improve the interpretability of our best-performing model, we included SHapley Additive exPlanations (SHAP) values with the Local Interpretable Model-Agnostic Explanations (LIME) technique. By quantifying each feature's contribution, SHAP values shed light on how these factors affect the best prediction model's results. Concurrently, the LIME technique helps assess how explainable the machine learning model is for predictions about specific patients.

The technique of Shapley Additive Explanation (SHAP) is utilised to provide explanations for the predictions generated by machine learning models. In consideration of all possible feature combinations, a game-theoretical approach allocates a value to each input feature to represent its contribution to the prediction. Healthcare, computer vision, and natural language processing have all implemented the SHAP algorithm [42]. Research has demonstrated that it offers explanations for complex machine learning model predictions that are both more precise and easily understood than alternative methodologies.

The formula for the SHAP algorithm is:

$$g(z) = \varphi_0 + \sum_{i=1}^{M} \varphi_i z_i \tag{10}$$

In the case where p represents the number of features and an input $[\times 1, \times 2, \dots, xp]$ is accompanied by a trained model f, SHAP approximates f using a straightforward model g. This approximation enables the determination of the effect of each feature on the prediction for every possible subset of features by accounting for the contribution of each feature value. The expression $[z1, z2, \dots; zp]$ simplifies the input x by assigning a value of 1 to z, which represents the features utilised in the data prediction process; a value of 0 indicates that the corresponding feature is not incorporated. $\phi i \in R$ denotes the Shapley value for each feature, which is calculated as the weighted sum of the contributions from all feasible subsets. The weights assigned to the features are proportional to the number of features present in each subset.

The autonomous instance predictions of machine learning models are elucidated using the interpretable machine learning framework LIME. LIME examines the impact of modifying the feature values of a solitary data sample on the output. LIME generates a novel dataset consisting of permuted samples and their corresponding black box model predictions, in accordance with this concept. The creation of the dataset includes, among other methods, the concealment of image portions, the removal of words (for NLP problems), and the addition of noise to continuous features [43]. The mathematical expression for local surrogate models that satisfy the interpretability requirement is as follows:

$$Interpretation\ (x) = arg.min_{v \in V} L(u, v, \pi_x) + w(v) \tag{11}$$

It can be deduced that the LIME algorithm is executed in the following manner:

- Select a particular instance from the input data that will serve as the basis for elucidating the model's prediction.
- In order to perturb the chosen instance, a collection of marginally altered data points is generated in the vicinity of the original point.
- Utilise the black-box model to forecast the outcome for every perturbed data point while documenting the corresponding input features.
- One potential approach is to train a locally interpretable model, such as a decision tree or linear regression, utilising the documented input features and output values.
- Using the local model, describe the prediction of the black-box model on the original data point.
- To validate the predictions of sophisticated machine learning models, LIME has been applied to a variety of fields, including computer vision, natural language processing, and healthcare. It is possible to use it to ascertain which input features have the greatest impact on the model's output for a specific instance of input data.

5.1 Performance measures

An approach for assessing a machine learning model is called a performance evaluation technique. A trained model's predictions made on the testing dataset are evaluated in this process. In this last step of the procedure, a thorough split of performance for detecting mental illness is achieved by applying specific performance measurements to the results collected in the previous phase. As shown in Table 2, accuracy, precision, recall, and F1 score are the key performance indicators we used to identify the predictive model of mental illness; among all methods used up until this point is the suitable model.

Table 2 Description of metrics used in the study

Metrics	Description	Formula
Accuracy [35]	It represents the overall sum of cases for which the model was successful in its recognition attempts	$= \dfrac{\sum_{i=1}^{I} TP_{i+} \sum_{i=1}^{I} TN_i}{\sum_{i=1}^{I} TP_{i+} \sum_{i=1}^{I} TN_{i+} \sum_{i=1}^{I} FP_{i+} \sum_{i=1}^{I} FN_i} \tag{12}$
Precision [37]	It separates real occurrences from predicted ones	$= \dfrac{\sum_{i=1}^{I} TP_i}{\sum_{i=1}^{I} TP_i + \sum_{i=1}^{I} FN_i} \tag{13}$
Recall [38]	True positives are the fraction of real cases that were really discovered	$= \dfrac{\sum_{i=1}^{I} TP_i}{\sum_{i=1}^{I} TP_i + \sum_{i=1}^{I} FN_i} \tag{14}$
F1-measure [39]	The model is tested by taking the harmonic-mean of the two metrics, Precision and Recall	$= \dfrac{\sum_{i=1}^{I} TP_i}{\sum_{i=1}^{I} TP_i + \sum_{i=1}^{I} FN_i} \tag{15}$

6 Experimental analysis and discussion

The concept of explainability becomes relevant when evaluating the accuracy of predictions made by each classification technique during a classification evaluation. First, calculate the ratio of accurate forecasts to inaccurate predictions. Additionally, a classification report's metrics are predicted depending on the pattern of true positive, false positive, false negative and true negatives. Accuracy is a metric that may be used to determine the efficacy of categorization systems. The proportion of correct predictions our model makes in a broad sense is how well it performs. A data item that algorithm correctly identified as true or false is known as a true positive or true negative value in computing. On the other side, a data item mistakenly identified by the algorithm and subsequently labeled as such is a false positive or false negative. The True Positive Rate is another term for sensitivity (TPR). Understanding the model's performance is essential when an observation falls under a different category than the studied one. The denominator rises, and the accuracy falls when the model produces many wrong positive classifications or few correct positive classifications. The prerequisites for model training, including the crucial tools and libraries, are shown in Table 3. Table 4 displays the total accuracy for each of the five ML techniques with various scaling approaches. Compared to the other four algorithms, the Gaussian & Random Forest algorithms demonstrated the best accuracy (min–max scaler) of 91% and 90%, respectively, when used without scaling approaches. In contrast, Linear Regression had a minor performance with 50 percent accuracy. The study's findings also demonstrated that the overall performance of all algorithms, apart from decision tree and linear regression, is comparable with or without data scaling strategies.

Figure 6a, b, and c show various performance scores of precision, recall, and F1 Scores. K-NN and decision tree on both scaling methods record the most remarkable

Table 3 Prerequisites for model training

Implementation Requirements	
Processor	Intel i7(Recommended)
Ram	32 GB
GPU	4 GB
Windows	10
Software	https://www.anaconda.com/
Web-based interactive environment	Jupyter Notebook
Programming language	Python

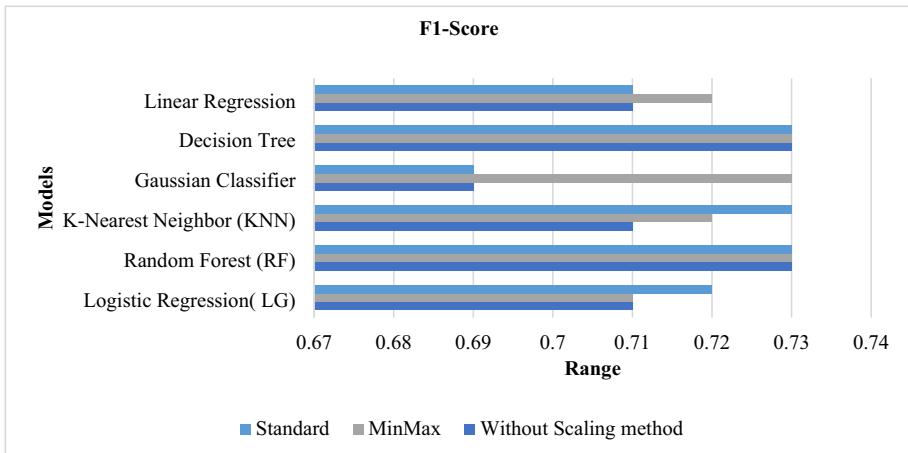**Table 4** Performance of various models based on Accuracy

Model	Accuracy (without any Scaling method)	MinMax Scaler	Standard Scaler
Logistic Regression (LG)	0.88	0.91	0.88
Random Forest (RF)	0.88	0.9	0.86
K-Nearest Neighbor (KNN)	0.89	0.87	0.82
Gaussian Classifier	0.86	0.91	0.90
Decision Tree	0.65	0.73	0.75
Linear Regression	0.5	0.89	0.84
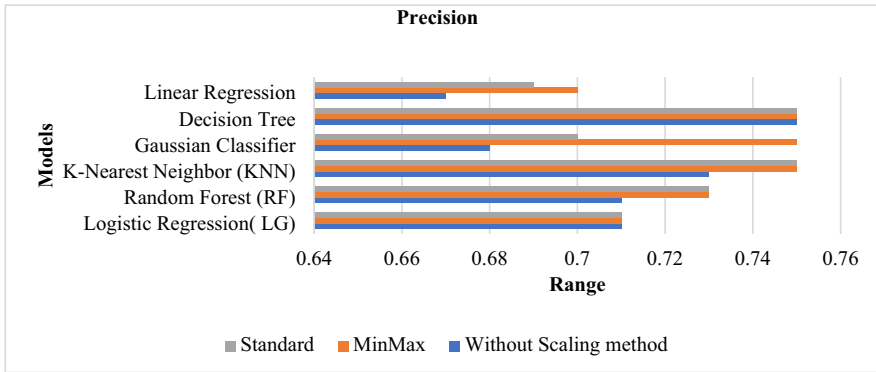
a
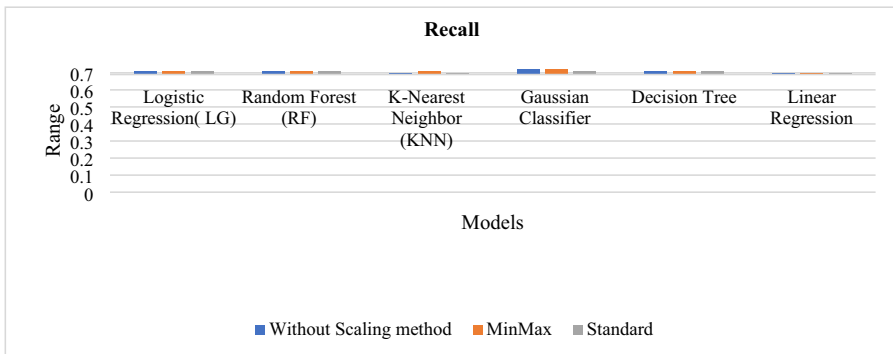
b

c

Fig. 6 **a** Results of Precision (Standard Scaler, Minmax Scaler and without any Scaling method), **b** Results of Recall (Standard Scaler, Minmax Scaler and without any Scaling method), **c** Results of F1-Score (Standard Scaler, Minmax Scaler and without any Scaling method)

results in terms of accuracy (standard scaler and min max scaler). Regarding recall, the gaussian classifier fared better on the min–max scaler. The F1-score results showed that random forest and Gaussian classifiers had the most significant scores.

6.1 Local Interpretation with SHAP and LIME

The global feature importance is depicted in Fig. 7 as the mean of the absolute SHAP values for every input variable. An increased mean SHAP value is indicative of a more pronounced influence. Mentally ill and age are the two most significant variables. The impact of each factor on the prognosis is represented by the arrows in Fig. 7; the colour blue and red signifies whether the factor had a decreasing or increasing effect on the prediction, respectively. The basal value is the mean of the database's predictions; the length of the bar represents the magnitude of the associated increases and decreases.as illustrated in Fig. 7.

The LIME local interpretation technique is utilised in Fig. 7b to provide a concise overview of the elements that contribute to the prediction. The factors designated with an orange colour signifies a favourable impact on the prediction of productivity, whereas the factors designated in blue represent an unfavourable contribution. The table located to the figure's right presents the actual value of each feature in addition to its ranking in terms of contribution to the prediction.

The findings indicate that age > 60, income, and age 45–60 all have a positive impact on the predicted productivity. This aligns with the SHAP local interpretation. Conversely, gender, unemployment, mental illness, and age 30–44 all exert an adverse influence on the prediction. The integration of LIME and SHAP yields a more comprehensive and detailed comprehension of the model's decision-making process, thereby potentially enhancing the model's efficacy and fostering confidence among petroleum engineers.
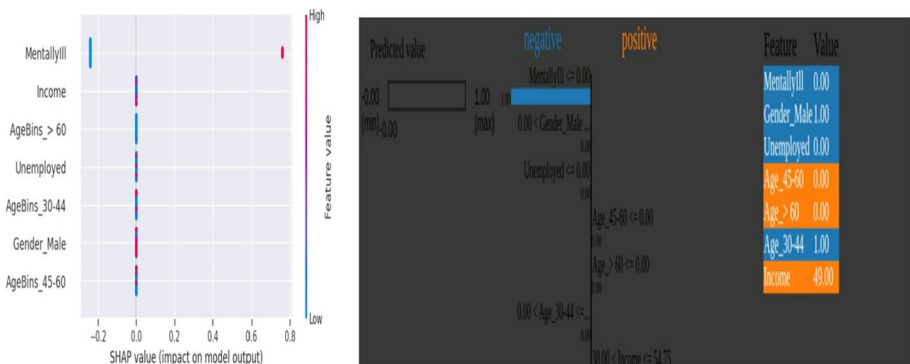

Fig. 7 Local interpretations **a**) with SHAP values **b**) with LIME

7 Conclusion and future directions

This research assessed six ML methods and two different data scaling approaches for identifying individuals with mental illness using the "mental health prediction" dataset from Kaggle Our results imply that data scaling techniques influence ML predictions in some way, contributing to the explainability of the models. The gaussian classifier beat all other methods suggested in this research, with more than 90% accuracy. This research also assessed performance variation while considering various data scaling techniques. Results reveal that algorithm performance varies with various scaling techniques. By utilising the SHAP-based model interpretability technique, the predicted outcomes were rendered in a comprehensible manner with regard to the relative significance of various input features. The examination of the SHAP values revealed that the well productivity is most significantly influenced by age and income. Additionally, the evaluation of feature contribution to the prediction of productivity for a solitary fractured well using SHAP and LIME yielded a comprehensive comprehension of the decision-making mechanism employed by the model. The application of local interpretation techniques proved to be advantageous in the detection of possible errors, biases, or opportunities for enhancement in the model. This is critical for fostering confidence among engineers and may potentially contribute to the advancement of more accurate physics-based predictive models specifically designed for undrilled wells. The findings may help early adopters find an optimum approach due to the quick progress in bioinformatics and other medical-based industries. As a result, ML models may have a significant influence not just on economics but also on helping healthcare practitioners by offering crucial insights. Such an algorithm can be merged into a chip and can be used to automate human–machine interaction. In future, results can be computed with more ML methods utilizing real-time mental disease data, modifying the settings of various data scaling approaches. Furthermore, future research will concentrate on deep learning approaches that will be examined to show how well such models performed in various applications.

Declarations

References

1. Baumann AE (2007) Stigmatization, social distance and exclusion because of mental illness: the individual with mental illness as a 'stranger.' Int Rev Psychiatry 19(2):131–135
2. Silvana M, Akbar R, Audina M (2018) Development of classification features of mental disorder characteristics using the fuzzy logic Mamdani method. In 2018 International Conference on Information Technology Systems and Innovation (ICITSI) pp 410–414. https://ieeexplore.ieee.org/abstract/document/8696043/
3. Silva C, Saraee M, Saraee M (2019) Data science in public mental health: a new analytic framework. In 2019 IEEE Symposium on Computers and Communications (ISCC) pp 1123–1128. https://ieeexplore.ieee.org/abstract/document/8969723

4. Gore E, Rathi S (2019) Surveying machine learning algorithms on EEG signals data for mental health assessment. In 2019 IEEE Pune Section International Conference (PuneCon) pp 1–6. https://ieeexplore.ieee.org/abstract/document/9105749

5. Binder MR (2021) The neuronal excitability spectrum: A new paradigm in the diagnosis, treatment, and prevention of mental illness and its relation to chronic disease. Am J Clin Experiment Med 9(6):187–203

6. Bailey F, Eaton J, Jidda M, van Brakel WH, Addiss DG, Molyneux DH (2019) Neglected tropical diseases and mental health: progress, partnerships, and integration. Trends Parasitol 35(1):23–31

7. Low DM, Bentley KH, Ghosh SS (2020) Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope Investigative Otolaryngology 5(1):96–116

8. Liang Y, Zheng X, Zeng DD (2019) A survey on big data-driven digital phenotyping of mental health. Information Fusion 52:290–307

9. Sau A, Bhakta I (2019) Screening of anxiety and depression among the seafarers using machine learning technology. Informat Med Unlocked 16:100149

10. Braithwaite SR, Giraud-Carrier C, West J, Barnes MD, Hanson CL (2016) Validating machine learning algorithms for Twitter data against established measures of suicidality. JMIR mental health 3(2):e4822

11. Srividya M, Mohanavalli S, Bhalaji N (2018) Behavioral modeling for mental health using machine learning algorithms. J Med Syst 42(5):1–12

12. Watts D, Moulden H, Mamak M, Upfold C, Chaimowitz G, Kapczinski F (2021) Predicting offenses among individuals with psychiatric disorders-A machine learning approach. J Psychiatr Res 138:146–154

13. Hornstein S, Forman-Hoffman V, Nazander A, Ranta K, Hilbert K (2021) Predicting therapy outcome in a digital mental health intervention for depression and anxiety: A machine learning approach. Digital Health 7:20552076211060660

14. Jain T, Jain A, Hada PS, Kumar H, Verma VK, Patni A (2021) Machine learning techniques for prediction of mental health. In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) pp 1606–1613. https://ieeexplore.ieee.org/abstract/document/9545061

15. Andersson S, Bathula DR, Iliadis SI, Walter M, Skalkidou A (2021) Predicting women with depressive symptoms postpartum with machine learning methods. Sci Rep 11(1):1–15

16. Sutter B, Chiong R, Budhi GS, Dhakal S (2021) Predicting psychological distress from ecological factors: a machine learning approach. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. pp 341–352. https://link.springer.com/chapter/10.1007/978-3-030-79457-6_30

17. Rahman JS, Gedeon T, Caldwell S, Jones R, Jin Z (2021) Towards effective music therapy for mental health care using machine learning tools: human affective reasoning and music genres. J Artif Intell Soft Comput Res. https://sciendo.com/article/10.2478/jaiscr-2021-0001

18. Sau A, Bhakta I (2017) Predicting anxiety and depression in elderly patients using machine learning technology. Healthcare Technol Lett 4(6):238–243

19. Sano A, Taylor S, McHill AW, Phillips AJ, Barger LK, Klerman E, Picard R (2018) Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: observational study. J Med Internet Res 20(6):e9410

20. Zulfiker MS, Kabir N, Biswas AA, Nazneen T, Uddin MS (2021) An in-depth analysis of machine learning approaches to predict depression. Current Res Behav Sci 2:100044

21. Edgcomb JB, Thiruvalluru R, Pathak J, Brooks JO III (2021) Machine learning to differentiate risk of suicide attempt and self-harm after general medical hospitalization of women with mental illness. Med Care 59:S58–S64

22. Kourou K, Manikis G, Poikonen-Saksela P, Mazzocco K, Pat-Horenczyk R, Sousa B, Fotiadis DI (2021) A machine learning-based pipeline for modeling medical, socio-demographic, lifestyle and self-reported psychological traits as predictors of mental health outcomes after breast cancer diagnosis: An initial effort to define resilience effects. Comp Biol Med 131:104266

23. Saba T, Khan AR, Abunadi I, Bahaj SA, Ali H, Alruwaythi M (2022) Arabic speech analysis for classification and prediction of mental illness due to depression using deep learning. Comput Intell Neurosc. https://www.hindawi.com/journals/cin/2022/8622022/

24. Linardon J, Fuller-Tyszkiewicz M, Shatte A, Greenwood CJ (2022) An exploratory application of machine learning methods to optimize prediction of responsiveness to digital interventions for eating disorder symptoms. Int J Eating Disord. https://onlinelibrary.wiley.com/doi/full/10.1002/eat.23733

25. Kumar S, Chong I (2018) Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states. Int J Environ Res Public Health 15(12):2907

26. Bartlett CL, Glatt SJ, Bichindaritz I (2019) Machine learning and feature selection for the classification of mental disorders from methylation data. In Conference on Artificial Intelligence in Medicine in Europe, Springer, Cham, pp 311–321.https://link.springer.com/chapter/10.1007/978-3-030-21642-9_40

27. Diaz-Ramos RE, Gomez-Cravioto DA, Trejo LA, López CF, Medina-Pérez MA (2021) Towards a resilience to stress index based on physiological response: A machine learning approach. Sensors 21(24):8293
28. Jaworska N, De la Salle S, Ibrahim MH, Blier P, Knott V (2019) Leveraging machine learning approaches for predicting antidepressant treatment response using electroencephalography (EEG) and clinical data. Front Psych 9:768
29. Kumar P, Chauhan R, Stephan T, Shankar A, Thakur S (2021) A machine learning implementation for mental health care. Application: Smart watch for depression detection. In 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) pp 568–574. https://ieeexplore.ieee.org/abstract/document/9377199
30. Prakash A, Agarwal K, Shekhar S, Mutreja T, Chakraborty PS (2021) An ensemble learning approach for the detection of depression and mental illness over twitter data. In 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom) pp 565–570. https://ieeexplore.ieee.org/abstract/document/9441288
31. Mishra S, Tripathy HK, Thakkar HK, Garg D, Kotecha K, Pandya S (2021) An explainable intelligence driven query prioritization using balanced decision tree approach for multi-level psychological disorders assessment. Front Publ Health. https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2021.795007/full
32. Alabi EO, Adeniji OD, Awoyelu TM, Fasae OD (2021) Hybridization of machine learning techniques in predicting mental disorder. Int J Human Computing Stud 3(6):22–30
33. Jan Z, Noor AA, Mousa O, Abd-Alrazaq A, Ahmed A, Alam T, Househ M (2021) The role of machine learning in diagnosing bipolar disorder: Scoping review. J Med Internet Res 23(11):e29749
34. Kim J, Lee D, Park E (2021) Machine learning for mental health in social media: bibliometric study. J Med Internet Res 23(3):e24870
35. Espinola CW, Gomes JC, Pereira JMS, dos Santos WP (2021) Vocal acoustic analysis and machine learning for the identification of schizophrenia. Res Biomed Eng 37(1):33–46
36. A Solanki, S Kumar, C Rohan, SP Singh, A Tayal (2021) Prediction of breast and lung cancer, comparative review and analysis using machine learning techniques. Smart Comput Self-Adapt Syst https://www.taylorfrancis.com/chapters/edit/10.1201/9781003156123-13/prediction-breast-lung-cancer-comparative-review-analysis-using-machine-learning-techniques-arun-solanki-sandeep-kumar-rohan-simar-preet-singh-akash-tayal
37. Elujide I, Fashoto SG, Fashoto B, Mbunge E, Folorunso SO, Olamijuwon JO (2021) Application of deep and machine learning techniques for multi-label classification performance on psychotic disorder diseases. Inform Med Unlocked 23:100545
38. Chahar R, Dubey AK, Narang SK (2021) A review and meta-analysis of machine intelligence approaches for mental health issues and depression detection. Int J Adv Technol Eng Explor 8(83):1279
39. Sahlan F, Hamidi F, Misrat MZ, Adli MH, Wani S, Gulzar Y (2021) Prediction of mental health among University Students. Int J Perceptive Cognitive Comput 7(1):85–91
40. Rana S, Soni V, Bairwa AK, Joshi S (2021) A review for prediction of anxiety disorders in humans using machine learning. In 2021 IEEE Bombay Section Signature Conference (IBSSC) pp 1–6. https://ieeexplore.ieee.org/abstract/document/9673471
41. Rainchwar P, Wattamwar S, Mate R, Sahasrabudhe C, Naik V (2021) Machine learning-based psychology: A study to understand cognitive decision-making. In International Advanced Computing Conference, Springer, Cham, pp 179–192. https://link.springer.com/chapter/10.1007/978-3-030-95502-1_14
42. Uddin MZ, Dysthe KK, Følstad A, Brandtzaeg PB (2022) Deep learning for prediction of depressive symptoms in a large textual dataset. Neural Comput Appl 34(1):721–744
43. Hosseinzadeh Kasani P, Lee JE, Park C, Yun CH, Jang JW, Lee SA (2023) Evaluation of nutritional status and clinical depression classification using an explainable machine learning method. Front Nutr 10:1165854

Authors and Affiliations

Inderpreet Kaur[1] · Kamini[1,3] · Jaskirat Kaur[1,4] · Gagandeep[1] · Simar Preet Singh[2] ⬤ **· Umesh Gupta[2]**

✉ Simar Preet Singh
  dr.simarpreetsingh@gmail.com; simarpreet.singh@bennett.edu.in

  Inderpreet Kaur
  cecm.bca.ik@gmail.com

  Kamini
  kamini.sehmi@sait.ca

  Jaskirat Kaur
  jaskiratkaur17@gmail.com; jaskiratkaur@pec.edu.in

  Gagandeep
  gaganpec@yahoo.com

  Umesh Gupta
  er.umeshgupta@gmail.com

[1] Chandigarh Group of Colleges, Landran, Mohali, India

[2] Bennett University, Greater Noida, Uttar Pradesh, India

[3] Southern Alberta Institute of Technology, Calgary, Alberta, Canada

[4] Punjab Engineering College (Deemed to be University), Chandigarh, India