Alexandria Engineering Journal 117 (2025) 376–390

Contents lists available at ScienceDirect

Alexandria Engineering Journal

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

Original Article

Explainable artificial intelligence systems for predicting mental health problems in autistics

El-Sayed Atlam [a,b,c,*], M. Rokaya [d,b], M. Masud [e], H. Meshref [e], Rakan Alotaibi [e], Abdulqader M. Almars [a], Mohammed Assiri [f], Ibrahim Gad [b]

[a] Department of Computer Science, College of Computer Science and Engineering, Taibah University, Yanbu 966144, Saudi Arabia
[b] Department of Computer Science, Tanta University, Tanta 31527, Egypt
[c] King Salman Center for Disability Research, Ola Abusukkar, 60-100, Riyadh, Saudi Arabia
[d] Department of Information Technology, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia
[e] Department of Computer Science, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia
[f] Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia

ARTICLE INFO

Dataset link: https://www.kaggle.com/code/monicabackes/mental-disorders-data-analysis/input

Keywords:
XAI
Machine learning
Mental health problems
Autistic
Disability
Analysis of Variance (ANOVA)
Mutual information (Mutinfo)

ABSTRACT

The recognition of mental disorder symptoms is crucial for timely management and reduction of recurring symptoms and disabilities. The ability to predict and explain mental health challenges can enable earlier intervention and more effective, individualized care plans, improving the overall well-being of people with autism. Consequently, artificial intelligence (AI) methods have been applied to assist psychologists and psychiatrists in decision-making processes by analyzing patients' medical histories and behavioral data. The current models for diagnosing mental health disorders (MHD) suffer from a lack of interpretability. This study introduces the **E**xplainable **M**ental **H**ealth **D**isorders **(EMHD)** model, a robust framework that leverages machine learning algorithms and Explainable Artificial Intelligence (XAI) to identify mental health disorders in young children, including toddlers. The EMHD consists of two main components: an ensemble model and Explainable Artificial Intelligence (XAI). First, an ensemble model known as Voting, which uses different feature selection techniques, namely Mutual Information (Mutinfo), Analysis of Variance (ANOVA) and Recursive Feature Elimination (RFE), is applied to classify the MHD dataset. Second, XAI is integrated into the proposed framework to provide transparency and explanations for the model's decision-making process. To achieve that, the model are explained using a well-known XAI technique called Shapley Additive Explanations (SHAP). The proposed EMHD demonstrates superior performance across all evaluation metrics, achieving an accuracy, precision, recall, and F1-Score of 1.0, in comparison to other baseline models. Furthermore, the study highlights the potential of XAI to provide personalized and actionable insights to mental health professionals who work with autistic individuals. Finally, this study can address the pressing MHD crisis in Saudi Arabia and significantly improve early MHD diagnosis.

1. Introduction

Mental health constitutes emotional, social, and psychological well-being of people, and can be adversely affected by various mental health disorders. People who suffer from these disorders can suffer from cognitive impairment, emotional distress, and interpersonal problems. To effectively manage these issues, it is crucial to perform comprehensive and timely assessments to distinguish and diagnose each disorder [1].

Based on information provided by the World Health Organization (WHO) [2], approximately 970 million adults worldwide, or one in every eight adults, experienced a mental health disorder in 2019. This figure highlights how common these diseases are in today's culture. In addition, the National Institute of Mental Health (2021) estimates that 22.8% of Americans experience mental illnesses. The contemporary way of living has a substantial impact on people's mental health and frequently causes emotional discomfort and depression [3,4]. A common psychiatric illness, depression impairs mental growth and cognitive abilities.

Addressing the magnitude of mental health issues requires a comprehensive approach [5,6]; World Health Organization, 2023). Despite the widespread prevalence of mental health disorders, current diagnostic and treatment methods exhibit several deficiencies. These include challenges in diagnosing due to the presence of comorbidities, which

* Corresponding author at: Department of Computer Science, College of Computer Science and Engineering, Taibah University, Yanbu 966144, Saudi Arabia.
E-mail address: stalam@taibahu.edu.sa (E.-S. Atlam).

https://doi.org/10.1016/j.aej.2024.12.120
Received 8 October 2024; Received in revised form 12 December 2024; Accepted 31 December 2024
Available online 16 January 2025

complicate the diagnostic process [7,8]. Clinicians often struggle with accurate diagnosis due to overlapping symptoms and uncertainties [7].

Furthermore, these difficulties are made worse by reliance on patients' erratic recollections of their activities and the scarcity of mental health specialists [9], making it more difficult to accurately diagnose medical issues and provide access to necessary therapy. In order to enhance the detection of mental health conditions, scholars have started examining the application of artificial intelligence models in literature reviews. These initiatives are motivated by AI's capacity to efficiently analyze large volumes of data [10], recognize unique patterns in data, comprehend important relationships between data points [11], and apply these understandings to generate precise predictions for new data [12–14].

The application of machine learning has yielded promising results for predicting and managing mental health disorders and other health-related conditions. These algorithms typically require extensive datasets to effectively learn patterns and perform classification tasks. While current studies demonstrate notable capability in identifying mental health disorders, they often lack explicit explanations for their findings. Hence, clinicians need to be able to interpret results in order to enhance their understanding of decision-making processes and improve diagnostic accuracy. The idea of XAI can help bridge the gap between the complex black boxes of AI models and the requirement for human understanding. In order to achieve this, XAI methods and tactics must be developed in order to increase comprehension and interpretation of AI models. This will enhance clinical understanding by improving explainability, encouraging meaningful interpretation and improving confidence in AI models [15,16] .

This study explores the application of XAI systems in predicting mental health problems in autistic patients. Given the unique challenges faced by this population, traditional predictive models often fall short in providing transparency and interpretability. In this paper, we present a novel model called Explainable Mental Health Disorders (EMHD), which utilizes machine learning (ensemble model) and XAI to identify mental health disorders in toddlers and children. This study's primary goal is to improve the comprehension and interpretation of the classification process while identifying the critical elements that influence MHD prediction. By integrating clinical data, behavioral assessments, and real-time monitoring, our approach seeks to empower healthcare providers with actionable information, ultimately improving patient outcomes and fostering personalized interventions. The suggested EMHD model shows remarkable performance in correctly classifying mental health illnesses, attaining excellent scores with accuracy, precision, recall, and an F1-Score of 1.0 across all evaluation parameters. Finally, this study can address the pressing MHD crisis in Saudi Arabia and significantly improve early MHD diagnosis. Experimental findings show that the model can detect MHDs and give reasons for its predictions.

The main contributions of this paper are:

- Development of a robust mental health disorder prediction model (EMHD).
- Integration of explainable AI (XAI) frameworks that enhance both predictive accuracy and understanding of the underlying causes of mental health issues.
- Implementation of SHAP (SHapley Additive exPlanations) for model interpretability, allowing for detailed explanations of individual predictions and identification of key contributing factors.
- Empirical validation through a comparative case study using publicly available datasets, demonstrating the model's feasibility.
- Superior performance compared to baseline models, achieving high accuracy and F1-scores in binary classification tasks.
- Improved early identification of mental health disorders, supporting healthcare professionals and psychiatrists.

2. Related work

The World Health Organization (WHO) states that in 2019, more than 970 million people around the world will have a mental health disorders. These numbers increased significantly once the COVID-19 pandemic struck in 2020, highlighting the critical need of addressing mental health issues and receiving appropriate medical treatment. Despite the availability of various treatment options, many individuals still face significant barriers in accessing these services. Notably, a considerable number of people experience discrimination, stigmatization, and violations of their human rights, further compounding their mental health challenges (World Health Organization) [17,18]).

A thorough mental health interview is usually necessary for the identification of mental health issues. A review of the patient's psychiatric history, physical examination, and reported symptoms are frequently included in this procedure. Psychometric tests and other evaluation instruments are also essential for detecting mental health issues [19].

Numerous studies have investigated the potential of machine learning (ML) in mental health, but also underscores critical limitations. For example, a scoping review highlighted the need for explainable ML models, noting that many current models lack transparency in their prediction processes. These models typically take raw data, such as MRI images, clinical notes, and EEG signals, as input and generate predictions without clear explanations of their inner workings [20]. Another study reviewed existing work on machine learning (ML) approaches for mental health diagnosis. It emphasized the importance of analyzing previous research to guide future directions. It found that ML models often outperformed standard ML models in terms of accuracy [21,22].

In a research study, Garcia-Ceja et al. [23] investigated the use of machine learning (ML) and sensor data for the diagnosis and treatment of mental health disorders. The study investigated several learning strategies, such as transfer learning, to treat psychological disorders like migraines, anxiety, and depression. It is crucial to remember, though, that the main contribution of this study was an overview of mental health problems and possible solutions.

In a different investigation, Gao et al. [23] examined brain imaging research using machine learning to predict illness [6,8,16,24]. The investigators examined bipolar disorder (BD) and major depressive disorder (MDD) using data from magnetic resonance imaging. The study assessed how well the decision tree (DT), logistic regression (LR), and support vector machine (SVM) algorithms predicted these mental health issues. In a different investigation, Gao et al. [23] examined brain imaging research using machine learning to predict illness [6,8, 16,24].

A follow-up study by Cho et al. [25] investigated five machine learning models for mental health issue detection: SVM, gradient boosting machines, random forests (RF), Naïve Bayes, and K-nearest neighbors (KNN). The detection of bipolar disorder, autistic spectrum disorder (ASD), schizophrenia, depression, and post-traumatic stress disorder (PTSD) was the main emphasis of the study. Predicting anxiety has proven to be difficult, in part because of its clinical parallels to MDD. The papers previously cited offer insightful information about the possible uses of ML in mental health diagnosis and research, which is worth emphasizing. To guarantee accuracy and dependability, it is crucial to take into account the constraints and current advancements in this discipline.

A study by Sau and Bhakta [26] used machine learning to predict symptoms of depression and anxiety in older adults. Furthermore, Gratch et al. [27] examined psychological distress disorders using interviews with patients. Yoon et al. [28] created a sophisticated multimodal DL model for the detection of depression. Xezonaki et al. [29] utilized a hierarchical attention network for classifying depression patient interviews.

ML models were used by Sharma et al. [30] on biomarker and self-reported depression data from the Lifelines Database. This study used

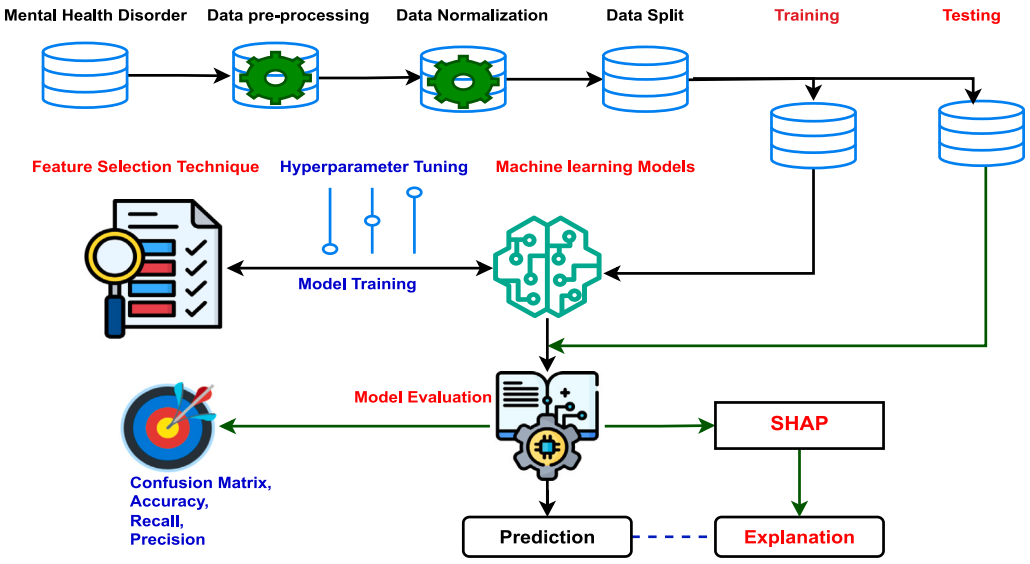

Fig. 1. The main steps of the Explainable Mental Health Disorders (EMHD) model.

many resampling strategies to solve the issue of uneven data within the dataset. The XGBoost algorithm was then applied to each individual sample for analysis. It is significant to highlight that these research offer insightful information on how ML might be used to predict and diagnose anxiety and depression. To improve the accuracy and dependability of the forecasts, correct data imbalances, and enhance these models, more study is needed.

A study examining the potential effects of artificial intelligence (AI) on mental health, including the assessment and reduction of algorithmic biases, was carried out by Timmons et al. [31]. They demanded that doctors be included in AI research, development, and care delivery, emphasizing the necessity of fair-aware AI in psychological study.

Rocheteau [32] utilized AI for diagnosing, monitoring, and prescribing therapies in mental health. Abdul Rahman et al. [33] conducted a broad cross-sectional study involving 17 Southeast Asian institutions, using machine learning techniques to predict mental health conditions. The study demonstrated that adaptive boosting algorithms and random forest (RF) may efficiently detect detrimental features of mental health. Furthermore, they emphasize the potential of AI in diagnosing and monitoring mental health conditions, while also underlining the need for ethical implementation and ensuring equitable access to mental healthcare.

An explainable model has been created by Atlam et al. [34] to detect autism spectrum disorder (ASD) in young children and toddlers. This model seeks to enhance our comprehension of the categorization process by identifying the important factors that influence the prediction of ASD. It is built on machine learning techniques. This study's main goal is to improve interpretability and shed light on the model's prediction-making process. Similarly, Mehedi et al. [35] have proposed a DL model for classifying individuals with Alzheimer's disease (AD) by utilizing magnetic resonance imaging (MRI) images. The goal of this research was to create a novel method that, using MRI data, can reliably identify AD patients.

3. Methodology

In this paper, we introduce a new model (EMHD) for predicting mental health problems in individuals with autism. Fig. 1 illustrates the main steps of the proposed EMHD Model. This flowchart depicts the various stages involved in developing the proposed model for mental health disorder classification. It highlights the importance of data preparation, model selection, interpretation, and evaluation in achieving accurate and reliable predictions. The proposed ensemble model, Voting, combines multiple feature selection methods, namely ANOVA, Mutinfo, and RFE, which are used for classifying the MHD dataset. The process can be summarized as follows:

3.1. Dataset gathering

This stage involves gathering and preparing the data for model training. The dataset on mental heath disorders (MHD) was gathered from the Kaggle repository and includes twelve different mental illnesses: Major Depressive Disorder (MDD), Obsessive-Compulsive Disorder (OCD), Bipolar Disorder, Eating Disorder, Sleeping Disorders, Psychotic Depression, Loneliness, Post-Traumatic Stress Disorder (PTSD), Autism Spectrum Disorder (ASD), and Pervasive Developmental Disorder (PDD).[1]

The dataset has 637 individuals. All columns except "Disorder" are integers, likely representing binary values (0 or 1) for the presence or absence of a symptom. "Disorder" is an object, it holds text labels for the different diagnoses. The dataset has no missing values. The column names are: 'age', 'trouble in concentration', 'feeling nervous', 'panic', 'hopelessness', 'sweating', 'close friend', 'trouble sleeping', 'feeling negative', 'trouble with work', 'breathing rapidly', 'over react', 'anger', 'change in eating','suicidal thought', 'feeling tired', 'seasonally', 'nightmares', 'weight gain','social media addiction', 'introvert', 'hallucinations', 'popping up stressful memory', 'blaming yourself', 'avoids people or activities', 'trouble concentrating', 'repetitive behavior', 'increased energy', and 'Disorder'. Further details on data statistics are provided in the Experimental Results and Analysis Section 4.

3.2. Preprocessing step

This stage involves preparing the data for model training, including data cleaning, normalization, and splitting into training and testing sets. At this stage, issues such as missing values, or inconsistencies in the collected data are handled. To ensure data integrity, we employ techniques such as data duplication, imputation of missing values, and elimination of inconsistencies. Furthermore, it is necessary to scale numerical features and encode categorical variables for this task.

A data preparation technique called feature selection is used to determine which subset of a dataset's columns are most pertinent to

[1] https://www.kaggle.com/code/monicabackes/mental-disorders-data-analysis/input Accessed date: May 15, 2024.

the target variable. By finding the most significant features, we can increase prediction models' performance and accuracy. Well-known method in feature selection is to assess each feature's association with the target variable using statistical metrics like mutual information. These measures assist in determining the significance of each feature to the predictive task. Feature selection algorithms can be customized to choose a specific number of attributes, varying from 1 to the whole number of attributes in the dataset.

Each set of chosen attributes is a subset of the input feature space, similar to a random subspace ensemble. However, the features are chosen based on a certain metric rather than being chosen randomly. To increase prediction performance, an ensemble model is developed with features chosen using feature selection techniques.

3.3. Normalization step

In machine learning, the normalization approach is frequently used to put a dataset's values on a consistent scale. This step ensures that all features are scaled to a common range, improving the model's training efficiency. Normalizing input features is performed to enhance the effectiveness of machine learning algorithms, which are often impacted by the volume of input data. This method entails rescaling an attribute's values to fit inside a range between 0 and 1. To achieve this, each individual value is subtracted by the minimum value of the attribute, and the result is divided by the range of the feature. This process ensures that the feature values are comparable and eliminates the bias that may be introduced by the original scale of the data as represented in Eq. (1):

$$\text{Normalized Value} = \frac{X - Mean}{\text{Standard Deviation}} \tag{1}$$

3.4. Data splitting

The primary objective of this phase is to divide the dataset into different subsets for the purposes of validation, testing, and training. These subsets can be generated using a variety of strategies, including stratified sampling and random splitting, to ensure their representativeness over the full dataset. The dataset is divided into two subsets: training and testing datasets. Usually, the split ratio used is around 80% for training data and 20% for testing data. The primary purpose of data splitting is to allow for more efficient model training and evaluation while minimizing the negative effects of overfitting and underfitting.

3.5. Machine learning models

This section outlines the feature extraction methods, the stacking ensemble machine learning models used by EMHD for predicting mental health conditions.

3.5.1. Feature extraction

The application of feature selection approaches to determine the most pertinent features for prediction is shown in Fig. 1. The aim of feature selection techniques is to determine the most informative and relevant features for a particular prediction task. The suggested model is capable of accurately capturing the underlying patterns and relationships seen in the data by choosing the best subset of attributes. Since the problem at hand is a classification task, three different feature selection techniques are commonly used: Mutual Information, ANOVA F-statistic, and RFE. Each method provides a unique perspective about the features most significant to the target variable. Feature selection methods can be implemented in several ways, but two popular approaches are as follows: a) One method: Create a feature subspace for each possible number of features, ranging from 1 to the total number of columns in the dataset. Train a model on each subset and merge their predictions. b) Multiple Methods: Create a feature subspace by employing several feature selection techniques. Train a model on each subset and merge their predictions.

3.5.2. Classification

This stage focuses on selecting and training the proposed machine learning model for the classification task. In machine learning, ensemble approaches can help to achieve this by combining the results of several basic models to increase prediction performance. An ensemble machine learning model called stacking combines the forecasts of several base models to produce predictions that are more accurate. It is particularly effective when many machine learning techniques exhibit skill on the same dataset, but in different ways. This indicates that the models' predictions or errors are unrelated or have a weak correlation. This approach effectively harnesses the collective knowledge of multiple models, providing a more robust and accurate prediction for classification or regression tasks. By combining these diverse predictions, stacking enables a meta-model to learn a more comprehensive representation of the data, potentially surpassing the performance of individual models.

In stacking, a diverse range of models is used as the base models, each making different assumptions about the prediction task. These base models collectively provide a more robust prediction by leveraging their individual strengths. To interpret the predictions made by the base models, a meta-model is used. The meta-model is typically straightforward, offering a seamless interpretation of the predictions made by the basic model. Linear models are widely utilized as meta-models in stacking. In the case of regression tasks, where the objective is to predict a numeric value, linear regression is commonly employed as the meta-model. On the other hand, for classification tasks (predicting a class label), logistic regression is widely used. However, it is important to note that while using linear models as the meta-model is common, it is not a strict requirement.

The prediction accuracy and generalization capacity of the trained model are assessed using the testing data that has not yet been observed. In this step, the model's performance is assessed using a variety of measures, including F1-score, recall, accuracy, and precision. Using the trained model to categorize new cases of mental health illnesses according to their textual characteristics is the last step.

3.5.3. Hyperparameter optimization

Grid search is a well-known technique for hyperparameter optimization, particularly effective when combined with ensemble methods in machine learning [36]. The core idea behind using grid search is to systematically explore a specified subset of hyperparameter values to identify the optimal combination to achieve optimal performance. In this paper, the key hyperparameters used by different models are :

- **n_estimators:** The number of trees or weak learners to be built in ensemble methods.
- **criterion:** Identifies the function (Gini Impurity) used to evaluate the quality of splits.
- **max_depth:** Controls the maximum depth of each tree. High number may leads to overfitting, so, Several optimal values are chosen to prevents trees from developing too deep. (See Section 4.3.1)
- **max_features:** Determines the number of features to evaluate when searching for the optimal split.
- **min_samples_split and min_samples_leaf:** Set the minimum number of data required to split a node and the minimum number of data needed at a leaf node.
- **Learning Rate (learning_rate):** applied in boosting algorithms like as Gradient Boosting and AdaBoost to understand how much each weak learner contributes to the final prediction.
- **splitter:** Provides the strategy for selecting the best split at each node (for decision trees).
- **Loss Functions (loss):** Indicates the model's optimization functions (e.g., m deviance, exponential).

The results are described in detail in Section 4.3.1.

3.6. Model interpretation (SHAP)

The XAI framework plays a vital role in identifying mental health disorders and providing meaningful interpretations of the model's outcomes. By leveraging advanced AI techniques, this framework enables the analysis and interpretation of complex data patterns associated with MHD. The framework uses the SHAP (SHapley Additive exPlanations) method at the interpretation step to accomplish this, giving insights into the importance of various criteria in the categorization process. In the training phase, the model picks up on relationships and patterns in the data that point to MHD. The procedure comprises adjusting the model's parameters to minimize the difference between the diagnoses for mental health disorders and the model's predictions. The XAI framework concentrates on ensuring interpretability for the model's predictions after it has been trained.

This is accomplished via a range of strategies meant to illuminate the model's decision-making procedure and the factors affecting its predictions. By employing the XAI framework, we not only identify mental health disorders but also comprehend how the model generates its forecasts. This interpretability improves our comprehension of the model's performance, gives the clinics trust in its results, and raises the system's general efficacy in the diagnosis of mental illnesses.

One interpretability method that is frequently utilized in XAI is feature importance analysis. Finding the factors that most significantly affect the model's predictions is its main goal. Researchers and doctors can obtain important insights into the elements most relevant for diagnosing mental health disorders by measuring the significance of each feature. The XAI framework uses visualization techniques in addition to feature importance analysis to give users a clearer grasp of the model's results. These visuals can be representations of the internal workings of the model, like decision trees that show the decision rules the model uses or heatmaps that show regions of interest in neuroimaging data.

In this work, we employ the force plot technique known as SHAP to visually understand individual predictions produced by the EMHD model. SHAP is a powerful technique used to explain the predictions made by machine learning models. By analyzing the contributions of each feature to the prediction outcome, SHAP offers insightful information on the model's decision-making procedure.

The SHAP values show how each feature is weighted differently in influencing the model's output. Features with larger bars have a greater effect on the result of the prediction than features with smaller bars. The purpose of the SHAP force plot is to provide a thorough explanation of the elements that go into a particular prediction for a particular instance. Along with their matching SHAP values, which show how much each characteristic contributes to the prediction, it shows the features that have an impact on the prediction. The average prediction for the complete dataset is also represented by a reference value on the SHAP force diagram. We present visualizations in Section 4 that illustrate how the SHAP force plot technique can be used to analyze the model's predictions. These visualizations offer a comprehensive understanding of the factors influencing the EMHD model's predictions and contribute to the overall interpretability of the system.

4. Experimental results and analysis

The following subsections present a detailed performance analysis and comparison with baseline methods.

4.1. Dataset description

In this section, we provide a detailed description of the MHD dataset utilized in this study. Table 1 presents a snapshot of symptoms and associated disorders, potentially collected from a survey or clinical records. Each row represents an individual, with columns indicating the presence or absence of various symptoms. The "age" column represents an identifier or age, while the remaining columns list symptoms like "feeling nervous", "panic", "breathing rapidly", and so on. The final column, "Disorder", labels the individual's diagnosed disorder, such as "MDD" (major depressive disorder), 'sleeping disorder', "Loneliness", 'PTSD', "ASD" (Autism Spectrum Disorder), 'ADHD', 'eating disorder', "bipolar", 'psychotic depression', 'PDD', 'OCD' and "anxiety". This data could be used for research purposes, such as identifying common symptom clusters associated with different disorders, or for developing diagnostic tools.

Table 1
A sample of rows of the Dataset.

	age	feeling nervous	panic	...	increased energy	Disorder
0	23	1	0	...	1	MDD
1	14	1	0	...	1	ASD
2	25	0	0	...	0	Loneliness
3	29	1	0	...	0	bipolar
4	32	1	1	...	1	anxiety

Fig. 2 visualizes the mean onset age of various mental health disorders with 95% confidence intervals. The x-axis displays the different disorders, including MDD, MHD (Mental Health Disorder), Loneliness, Bipolar Disorder, Anxiety, Sleep Disorders, Psychosis, Depression, Eating Disorders (ED), ADHD (Attention Deficit/Hyperactivity Disorder), PDD (Pervasive Developmental Disorder), and OCD (Obsessive-Compulsive Disorder). The y-axis represents the mean onset age in years. The Figure shows that the mean onset age varies significantly across different disorders. For example, the mean onset age for PTSD is significantly higher than that for MDD, ASD, or Loneliness. The error bars indicate the 95% confidence interval, suggesting the range within which the true mean onset age is likely to fall. This Figure provides a quick overview of the typical age at which different mental health disorders tend to emerge. It is crucial to note that these are only averages and that every person's experience will be very different.

Fig. 3 presents the incidences of various mental health symptoms. The x-axis lists different symptoms, such as "feeling nervous", "panic", "breathing rapidly", "sweating", "trouble concentrating", and so on. The y-axis represents the number of individuals experiencing each symptom. The figure shows that some symptoms are more prevalent than others. For example, "feeling nervous" and "breathing rapidly" have the highest incidences, while symptoms like "hallucinations" and "seasonally" have significantly lower incidences. This suggests that certain mental health symptoms are more common than others within the dataset. The figure provides a quick overview of the frequency of different mental health symptoms.

Fig. 4 displays the incidences of different diagnosed mental health disorders. The x-axis lists the disorders, including ASD, Anxiety, MDD, ED, ADHD, PDD, PTSD, Loneliness, Bipolar Disorder, Sleep Disorder, OCD, and Psychosis. The y-axis represents the number of individuals diagnosed with each disorder.

The chart shows that MHD has the highest incidence, followed by Anxiety, MDD, and ED. The incidences of other disorders decrease progressively, with Psychosis having the lowest incidence. The summary of the total incidences for each disorder: ASD: 87, Anxiety: 67, MDD: 66, ED: 65, ADHD: 65, PDD: 60, PTSD: 45, Loneliness: 44, Bipolar: 43, Sleep Disorder: 33, OCD: 29, and Psychosis: 26. This figure provides a quick overview of the relative prevalence of different mental health disorders within the dataset.

4.2. The performance matrices

The commonly used metrics, such as accuracy, precision, recall, and F1-Score, are used to evaluate the model's performance. Accuracy, as described by Eq. (2), evaluates the overall accuracy of the model's predictions by determining the ratio of correctly classified instances (True Positives (TP) and True Negatives (TN)) to the total number of instances (TP, TN, False Positives (FP), and False Negatives (FN)).

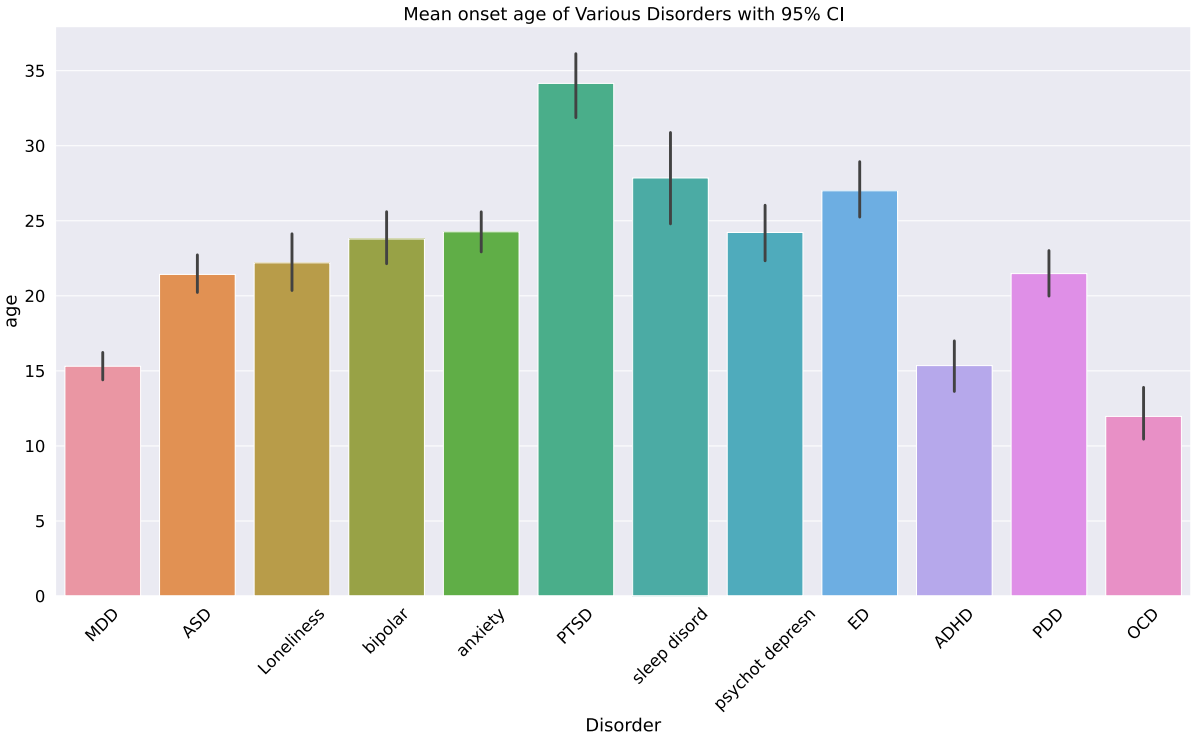

Fig. 2. Mean onset age of Various Disorders with 95% CI.

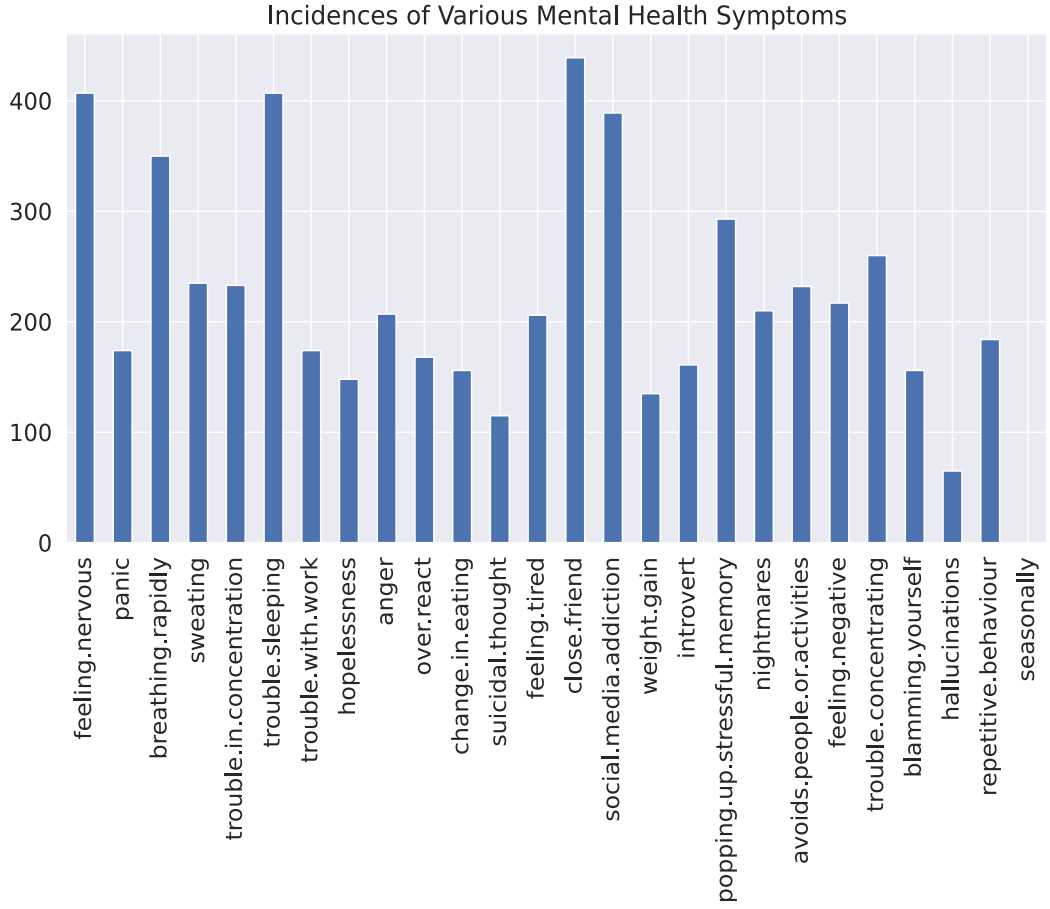

Fig. 3. Incidences of Various Mental Health Symptoms.

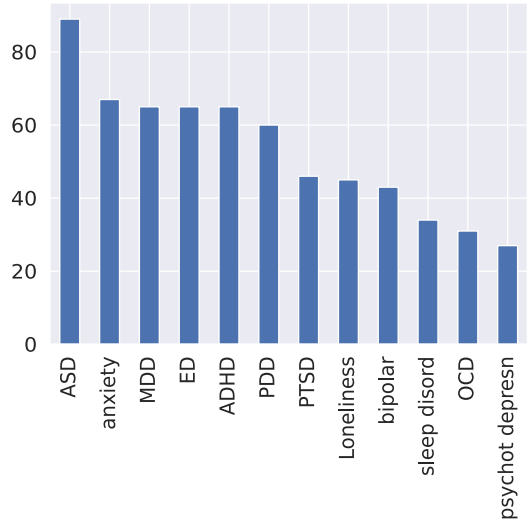

Fig. 4. The total incidences of each diagnosed disorder.

Where:

True Positive: An autistic individual with a history of anxiety is predicted to be at risk of developing depression, which is confirmed by subsequent clinical evaluation.

False Positive: An autistic individual without any significant mental health symptoms is predicted to be at risk, but further evaluation shows no mental health problems.

True Negative: An autistic individual with no mental health history is correctly predicted to be at low risk for mental health issues.

False Negative: An autistic individual at risk for developing depression is predicted to have no mental health concerns, but further clinical assessment reveals an emerging mental health problem. Then:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Precision, as defined by Eq. (3), measures the ability of the model to correctly identify positive instances. It is calculated as the ratio of True Positives to the sum of True Positives and False Positives.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Recall, defined in Eq. (4), determines the model's capability to correctly capture positive instances. It is calculated as the ratio of TP to the sum of TP and FN.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

The F1-Score, as given in Eq. (5), serves as a balanced measure between Precision and Recall, providing an aggregate assessment of the model's performance. It is computed as twice the product of Precision and Recall divided by their sum.

$$F1 - score = \frac{2(Precision * Recall)}{(Precision + Recall)} \tag{5}$$

4.3. Results and analysis

K-fold cross-validation is the most popular technique for creating the training dataset for the machine learning model. Using this method, the dataset is divided into k folds, or subsets, of roughly equal size. Every fold serves as a validation set during the model's k training and evaluation cycles, with the remaining folds being used for training.

The training dataset for the meta-model is created using the basic models' out-of-fold predictions. The training data for the meta-model may include the inputs utilized by the base models during their training, as well as the predictions generated by those models. This includes the input elements of the training data, providing valuable contextual information to the meta-model. By incorporating this information, the meta-model gains a deeper understanding of how to effectively combine and leverage the predictions from the base models. By separating the dataset into many folds and training the model on different subsets, we are able to evaluate the performance and generalizability of the meta-model more thoroughly. This robust training process, along with the inclusion of base model inputs, enhances the interpretability and predictive accuracy of the meta-model in the domain of mental health disorders.

4.3.1. Hyperparameters optimization using the grid search

The grid search yielded optimal hyperparameters for each of the five machine learning models evaluated. Table 2 presents the results of a grid search conducted to optimize the hyperparameters of five machine learning models for predicting mental health outcomes. The table includes the following information: Model: The name of the machine learning model. Parameters: The hyperparameters that were optimized during the grid search. Parameter values: The range of values considered for each hyperparameter. Best values: The optimal values for each hyperparameter that yielded the best performance. Accuracy: The overall accuracy of the model on the test data. Precision: the percentage of actual positive cases out of all anticipated positive cases. Recall: the percentage of accurately detected true positives. F1-score: The harmonic mean of precision and recall. Feature ranking: The top features identified as most important for prediction by the model. Importances: The relative importance scores assigned to each feature by the model.

This table offers a thorough summary of the performance and feature importance of each model under consideration. It facilitates a comparative analysis of the models and allows for the selection of the most suitable model for the specific task of mental health prediction. The best performing model was the Random Forest, achieving perfect accuracy, precision, recall, and F1 score. This model was configured with 50 estimators, the 'entropy' splitting criterion, a maximum depth of 7, and 4 maximum features.

For the Random Forest model, the best parameter values identified were 'n_estimators' = 50, 'criterion' = 'entropy', 'max_depth' = 7, 'max_features' = 4, "min_samples_split" = 2, and "min_samples_leaf" = 1. The corresponding feature ranking indicated that the most significant feature was 'breathing.rapidly' with an importance value of 0.0553, followed by 'hopelessness' (importance = 0.0425), 'trouble.in.concentration' (importance = 0.0388), 'anger' (importance = 0.0296), 'trouble.with.work' (importance = 0.0193), and 'age' (importance = 0.0182).

For the decision tree model, the best parameter values were 'criterion' = 'entropy', 'splitter' = 'best', and 'max_depth' = 6. The most relevant feature was 'trouble.sleeping' with an importance value of 0.0298, followed by 'sweating' (importance = 0.0038) and 'age' (importance = 0.0028). In the case of Gradient Boosting, the optimal parameter values were "loss" = 'deviance', "learning_rate" = 0.1, 'max_depth' = 4, 'min_samples_leaf' = 10, 'n_estimators' = 150, and 'max_features' = 0.3. The feature 'age' had the highest importance with a value of 0.0612, followed by 'trouble.in.concentration' (importance = 0.0456), 'breathing.rapidly' (importance = 0.0383), 'hopelessness' (importance = 0.0321), 'feeling.nervous' (importance = 0.0308), and 'trouble.sleeping' (importance = 0.0291).

For the AdaBoost model, the best values were 'learning_rate' = 0.1 and 'n_estimators' = 50. The most important features were 'feeling.nervous' (importance = 0.12) and 'hopelessness' (importance = 0.04). Lastly, for the Extra Trees model, the optimal parameter values were 'criterion' = 'gini', 'max_features' = 10, 'max_depth' = 7, and 'n_estimators' = 100. The most relevant features were 'hopelessness' (importance = 0.0664), 'trouble.sleeping' (importance = 0.0546), 'breathing.rapidly' (importance = 0.0442), and 'anger' (importance =

Table 2
The results of the ML models using Grid Search method.

Model	Parameters	Parameter values	Best values	Acc.	Prec.	Recall	F1	Feature ranking	Importances
Random Forest	'n_estimators'	[50, 100, 150, 200]	50	1	1	1	1	breathing.rapidly	0.0553
	'criterion'	['entropy', 'gini']	entropy					hopelessness	0.0425
	'max_depth'	[4, 6, 7]	7					trouble.in.concentration	0.0388
	'max_features'	[2, 3, 4]	4					anger	0.0296
	"min_samples_split"	[2, 3, 10]	2					trouble.with.work	0.0193
	"min_samples_leaf"	[1, 3, 10]	1					age	0.0182
Decision Tree	'criterion'	['entropy', 'gini']	entropy	1	1	1	1	trouble.sleeping	0.0298
	'splitter'	['best', 'random']	best					sweating	0.0038
	'max_depth'	[4, 5, 6, 7]	6					age	0.0028
Gradient Boosting	"loss"	["deviance", 'exponential']	deviance	1	1	1	1	age	0.0612
	"learning_rate"	[0.1, 0.05, 0.01]	0.1					trouble.in.concentration	0.0456
	'max_depth'	[4, 7]	4					breathing.rapidly	0.0383
	'min_samples_leaf'	[1,5, 10]	10					hopelessness	0.0321
	'n_estimators'	[50, 100, 150, 200]	150					feeling.nervous	0.0308
	'max_features'	[0.3, 0.1]	0.3					trouble.sleeping	0.0291
Extra Trees	'criterion'	['entropy', 'gini']	gini	1	1	1	1	hopelessness	0.0664
	"max_features"	[1, 3, 10]	10					trouble.sleeping	0.0546
	'max_depth'	[4, 5, 6, 7]	7					breathing.rapidly	0.0442
	'n_estimators'	[50, 100, 150, 200]	100					anger	0.0289
AdaBoost	'learning_rate':	[0.0001, 0.001, 0.01, 0.1, 0.2, 0.3,1.5]	0.1	0.99	0.99	0.99	0.99	feeling.nervous	0.1200
	'n_estimators':	[1, 2, 50, 100, 150, 200]	50					hopelessness	0.0400

0.0289).

Feature ranking analysis revealed that the most important features for predicting mental health outcomes were 'breathing.rapidly', 'hopelessness', 'trouble.in.concentration', 'anger', and 'trouble.with.work'. These features were consistently ranked highly across all models, suggesting their significant contribution to the prediction task.

4.3.2. Feature selection techniques

Table 3 presents the outcomes of the machine learning models employing various feature selection methods. The models' performance is evaluated using multiple metrics to assess their effectiveness. The decision tree (DT) model, serving as the baseline, demonstrates perfect performance across all metrics, including accuracy, precision, recall, and the F1 score. The subsequent Voting model, built on top of DT, also achieves identical performance without any feature selection. When incorporating feature selection techniques, such as RFE, ANOVA, and Mutinfo, the Voting model maintains its exceptional performance, with all metrics unchanged.

Ensemble approaches, such as Voting with ANOVA-mutinfo-RFE, Voting with many subsets of features, stacking with Logistic Regression (LR), DT, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Gaussian Naive Bayes (GaussianNB), and LR individually, exhibit consistent results across all metrics, achieving a high level of accuracy, precision, recall, and F1 score. On the other hand, the bagging model, constructed using DT as the base estimator, performs identically to the baseline DT model, indicating no improvement in performance. When ANOVA and mutual information methods are applied individually to the DT and LR models, the resulting models maintain high accuracy, precision, recall, and F1 score, similar to the baseline models.

Furthermore, the individual LR, GaussianNB, KNN, and SVC models achieve acceptable performance, although not as high as the DT-based models. These models exhibit comparatively lower accuracy, precision, recall, and F1 score. Similarly, LR, GaussianNB, and KNN models exhibited high performance with scores of 0.9922, 0.9688, and 0.9453, respectively. However, the SVC model displayed comparatively lower performance across all metrics, achieving scores of 0.53125 for accuracy, precision, recall, and F1 score. These results demonstrate the effectiveness of the proposed model and the feature selection techniques in achieving accurate predictions of the target variable. The models achieved high scores in the evaluation metrics, indicating their ability to capture relevant patterns and classify instances correctly. The proposed model, along with the selected feature selection techniques, can be considered as promising approaches for the given task.

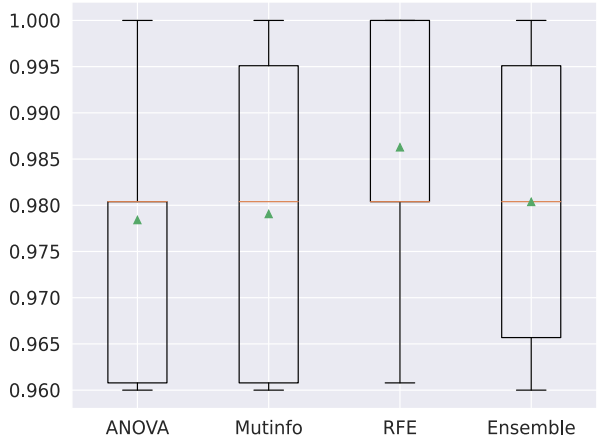

Fig. 5. Comparison of Voting Classifier of a fixed number attributes to single models fit on each set of features.

Fig. 5 visualizes the performance of four different feature selection methods: ANOVA, Mutinfo, RFE, and the voting method. The *y*-axis represents a performance metric accuracy, ranging from 0 to 1.0. Each boxplot represents the distribution of performance scores for a given method across multiple runs or trials. The box itself shows the interquartile range (IQR), with the middle line representing the median. The whiskers stretch up to 1.5 times the box's IQR; any data points outside of this range are represented by circles, which indicate outliers.

The summary of the performance based on the boxplots: 1) ANOVA: The median performance is around 0.981, with a relatively wide spread (IQR of 0.017). This suggests ANOVA might have some variability in its performance. 2) Mutinfo: Mutinfo has a slightly higher median performance (0.984) and a slightly narrower spread (IQR of 0.015). This indicates it is more consistent in its performance than ANOVA. 3) RFE: RFE shows the highest median performance (0.991) and the narrowest spread (IQR of 0.011). This suggests that RFE is the most reliable and consistent method among the four. 4) Ensemble: The ensemble method has a median performance of 0.986, with an IQR of 0.016. It shows a performance similar to Mutinfo, suggesting that combining multiple feature selection methods might provide a balanced performance.

Table 3
The results of the ML models using different feature selection methods.

Model	Baseline	Feature selection method			Metrics			
		RFE	ANOVA	Mutual	Accuracy	Precision	Recall	F1
Decision Tree (DT)	Baseline				1.0000	1.0000	1.0000	1.0000
Voting	DT	✓			1.0000	1.0000	1.0000	1.0000
The proposed model: Voting(ANOVA-Mutinfo-RFE)	DT	✓	✓	✓	1.0000	1.0000	1.0000	1.0000
DT		✓			1.0000	1.0000	1.0000	1.0000
Voting many subsets features	DT	✓	✓	✓	1.0000	1.0000	1.0000	1.0000
Stacking	LR, DT, KNN, SVC, GNB, LR				1.0000	1.0000	1.0000	1.0000
Bagging	DT				1.0000	1.0000	1.0000	1.0000
Voting	DT		✓		0.9922	0.9922	0.9922	0.9922
DT			✓		0.9922	0.9922	0.9922	0.9922
DT				✓	0.9922	0.9922	0.9922	0.9922
LR					0.9922	0.9922	0.9922	0.9922
GaussianNB					0.9688	0.9688	0.9688	0.9688
KNN					0.9453	0.9453	0.9453	0.9453
SVC					0.5313	0.5313	0.5313	0.5313

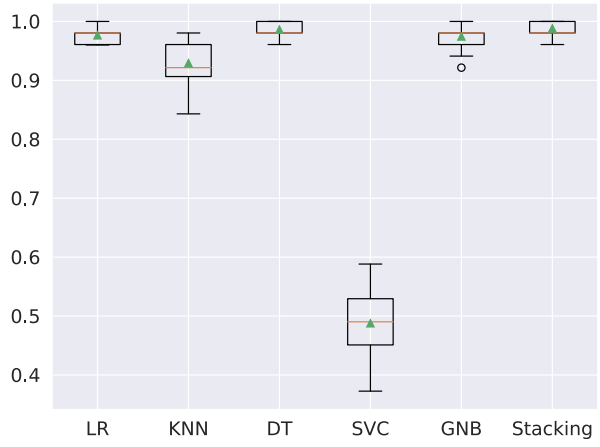

Fig. 6. Comparison of Stacking of a fixed number attributes to single models fit on each set of features.

Fig. 6 provides the performance of five different models: LR, KNN, DT, SVC, GNB, and the Stacking model. The performance metric accuracy is represented by the *y*-axis, which has a range of 0 to 1.0. The LR model achieves an accuracy of 0.983, with a standard deviation of 0.016. The KNN model attains an accuracy of 0.940, with a standard deviation of 0.031. The DT model demonstrates the highest performance among the models, with an accuracy of 0.991 and a standard deviation of 0.011. The SVC model shows a relatively lower accuracy of 0.547, with a standard deviation of 0.060. The GNB model obtains an accuracy of 0.973, with a standard deviation of 0.016. The Stacking model achieves the same accuracy as LR at 0.983, with a standard deviation of 0.016.

The confusion matrix of the SVC model is presented in Fig. 7. The confusion matrix provides an in-depth evaluation of the model's performance by illustrating the classification results for each class. Fig. 7 displays the distribution of predicted classes compared to the actual classes. The rows represent the actual classes, while the columns represent the predicted classes. The numerical values within the matrix indicate the count of instances assigned to each category.

The SVC model exhibited varying levels of accuracy in classifying instances across different classes. Some classes were correctly classified with a high degree of accuracy, as indicated by the large values along the diagonal of the confusion matrix. These instances were accurately assigned to their respective classes.

However, the model also encountered challenges in accurately classifying certain instances. This is evident from the off-diagonal values in the confusion matrix, which represent misclassifications. The model demonstrated lower accuracy in predicting instances for these specific classes, resulting in a higher number of false positives and false negatives.

The confusion matrix of the proposed model is presented in Fig. 8. Fig. 8 illustrates the distribution of predicted classes versus the actual classes. The rows represent the actual classes, while the columns depict the predicted classes. The numbers within the matrix indicate the count of instances falling into each category.

The proposed model demonstrated notable performance in classifying instances across different classes. It achieved a high number of correct predictions, particularly for classes with larger counts. The model accurately classified the majority of instances belonging to these classes, as evidenced by the high values along the diagonal of the confusion matrix.

4.3.3. The results of XAI

SHAP is a powerful technique used to explain the predictions made by machine learning models. By analyzing the contributions of each feature to the prediction outcome, SHAP provides valuable insights into the model's decision-making process. In Fig. 9, the SHAP values for the features are visualized, illustrating their impact on the model's predictions. Each feature is represented by a vertical bar, and the length and direction of the bar indicate the magnitude and direction of its influence on the prediction, respectively.

The SHAP values reveal the relative importance of each feature in determining the model's output. Features with longer bars have a stronger influence on the prediction outcome, while features with shorter bars have less impact. By examining the SHAP values, it is possible to detect the characteristics that contribute the most to the model's predictions. These attributes are critical in shaping the result and offer insightful information about the underlying relationships and patterns that the model captures.

Moreover, each feature's direction of influence can be determined with the aid of the SHAP values. Higher values of the appropriate feature positively contribute to the forecast when the SHAP values are positive; conversely, negative SHAP values suggest the reverse, as seen in Fig. 10.

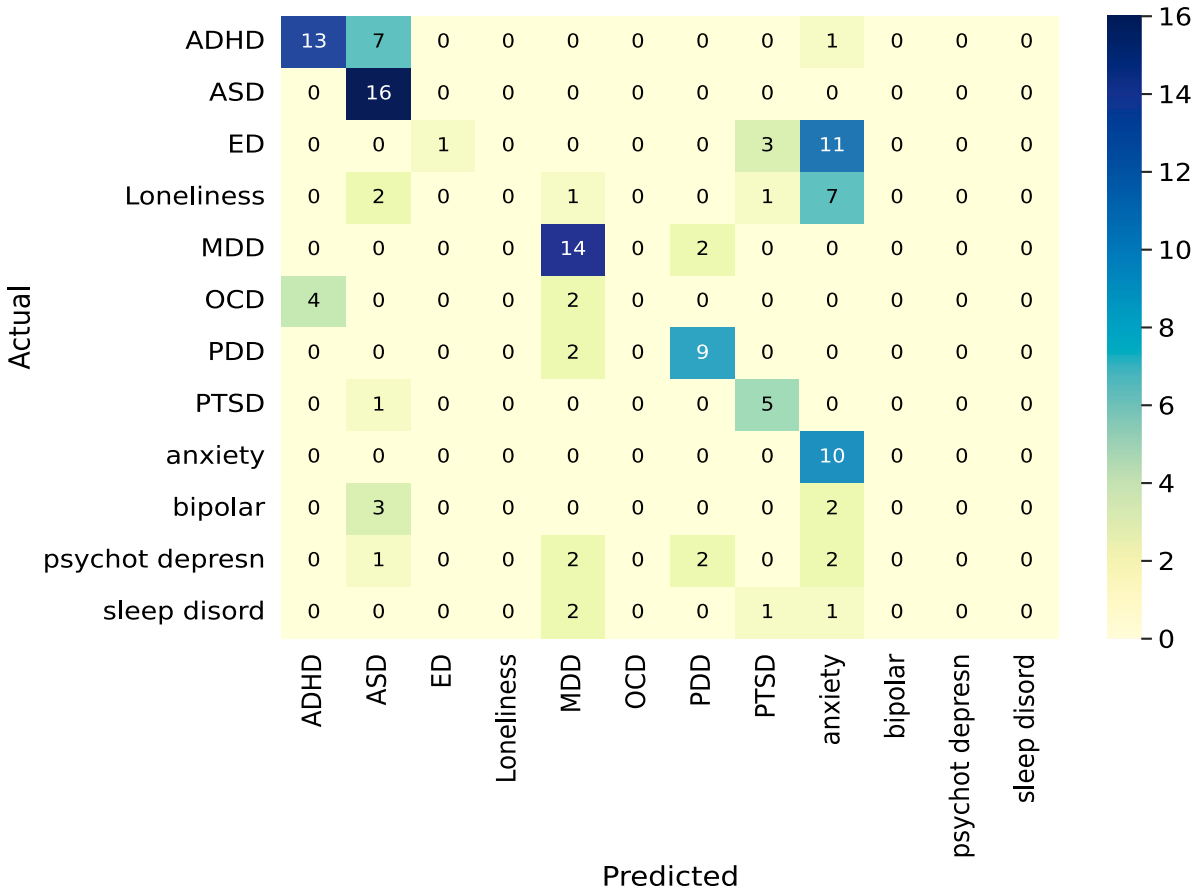

Fig. 7. The confusion matrix of SVC Classifier.

The SHAP approach offers a thorough comprehension of the ways in which each characteristic (feature) influences the predictions made by the model. The summary plot violin, shown in Fig. 11, provides a visual representation of the impact of various features on the model output. This plot offers insights into the relationship between each feature and the predicted outcome, allowing for a comprehensive analysis of their influence. Each feature is represented by a violin-shaped plot, where the width of the plot corresponds to the density of data points. The vertical axis represents the range of feature values, while the horizontal axis displays the distribution of SHAP values. The SHAP values represent the impact of each feature on the model output.

Fig. 11 depicts the distribution of SHAP values corresponding to each feature. Features that have larger absolute SHAP values have a stronger effect on the model's results than features that have lower absolute SHAP values. The width of the violin plot indicates the density of data points and provides an understanding of the variation in feature values. In this summary plot violin, features such as "feeling negative", "hopelessness", and "trouble sleeping" exhibit wider and more concentrated plots, suggesting their strong association with the model predictions. On the other hand, features like "age" and "increased energy" display narrower and more dispersed plots, indicating a relatively weaker influence on the model output.

The decision tree classifier's waterfall plot, shown in Fig. 12, illustrates how features affect the projected result one after another in a sequential fashion. This plot facilitates a thorough analysis of the contribution of each feature to the final prediction.

The waterfall plot begins with the initial value of the function, f(x), set to 0. It then proceeds to display the impact of each feature on the prediction. Each feature is represented by a bar, and the length and direction of the bar indicate the magnitude and direction of its influence. The bars are arranged in descending order based on their absolute impacts, with the most influential features appearing at the top. Features with negative values indicate those that lower the score, and those with positive values suggest those that raise the prediction score. The cumulative effect of the features is depicted by the cumulative sum of the bars.

In this waterfall plot, it can be observed that the feature "age" has the most significant positive impact on the prediction, contributing a value of 0.94. Other features such as "close.friend", "hopelessness", and "introvert" also have positive impacts, albeit with smaller values. On the other hand, features like "feeling.negative", "feeling.tired", and "sweating" have relatively smaller positive impacts. The feature "trouble.sleeping" has a neutral impact, indicated by a bar of zero length. The feature "nightmares" does not appear to contribute to the prediction. The cumulative sum of the impacts of these features results in the final prediction score, E[f(X)], which is determined to be 3.99 in this case.

5. Discussion

The proposed ensemble model, Voting with ANOVA, Mutinfo, and RFE, achieves perfect scores across all evaluation metrics. This indicates that the ensemble approach, combining the strengths of multiple feature selection techniques, results in a highly accurate and precise model for mental health disorders classification. With an accuracy, precision, recall, and F1-Score of 1.0, this ensemble model demonstrates exceptional capabilities in accurately classifying mental health disorders. Similarly, the Stacking model also achieves perfect scores across all evaluation metrics, further affirming its effectiveness in accurately classifying mental health disorders. These models demonstrate superior accuracy, precision, recall, and F1-score compared to the KNN and LSTM models, showcasing their potential as reliable tools for mental health disorders classification.

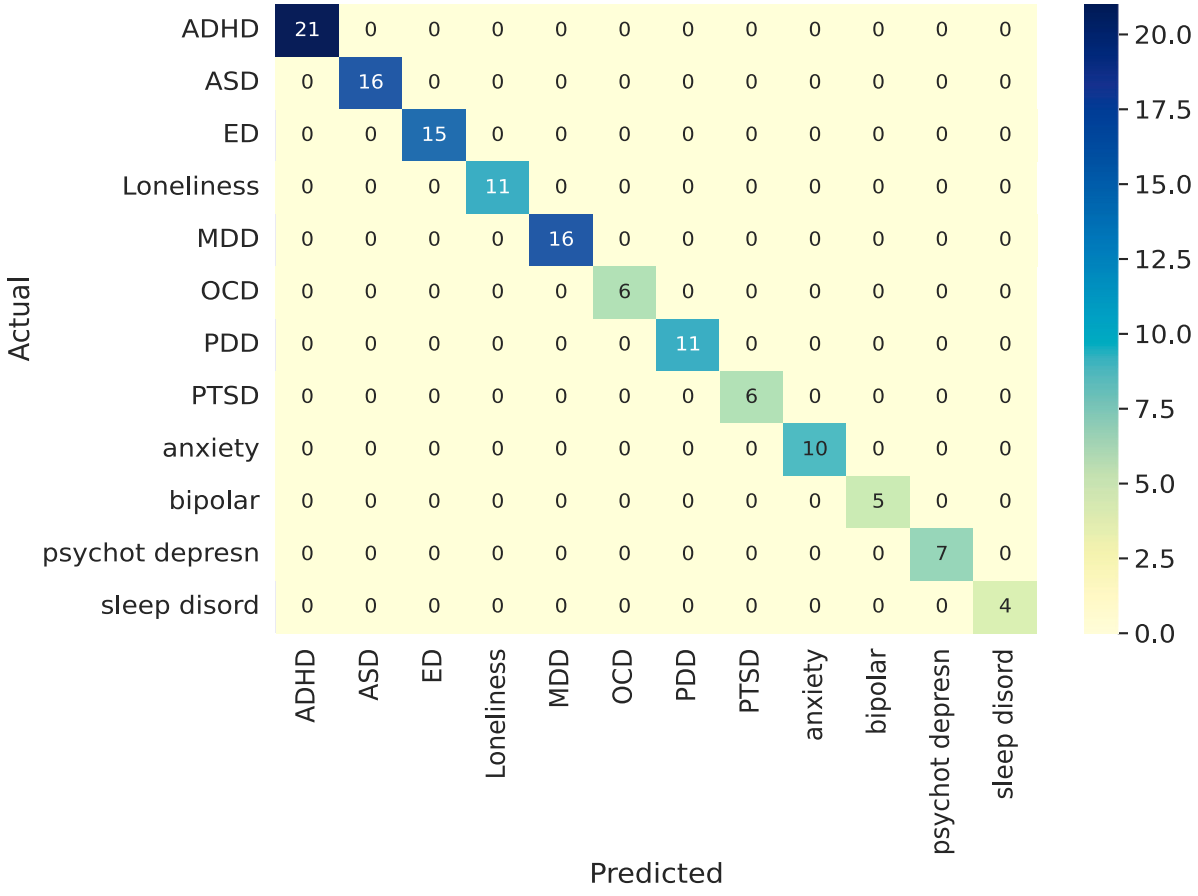

Fig. 8. The confusion matrix of the proposed model.

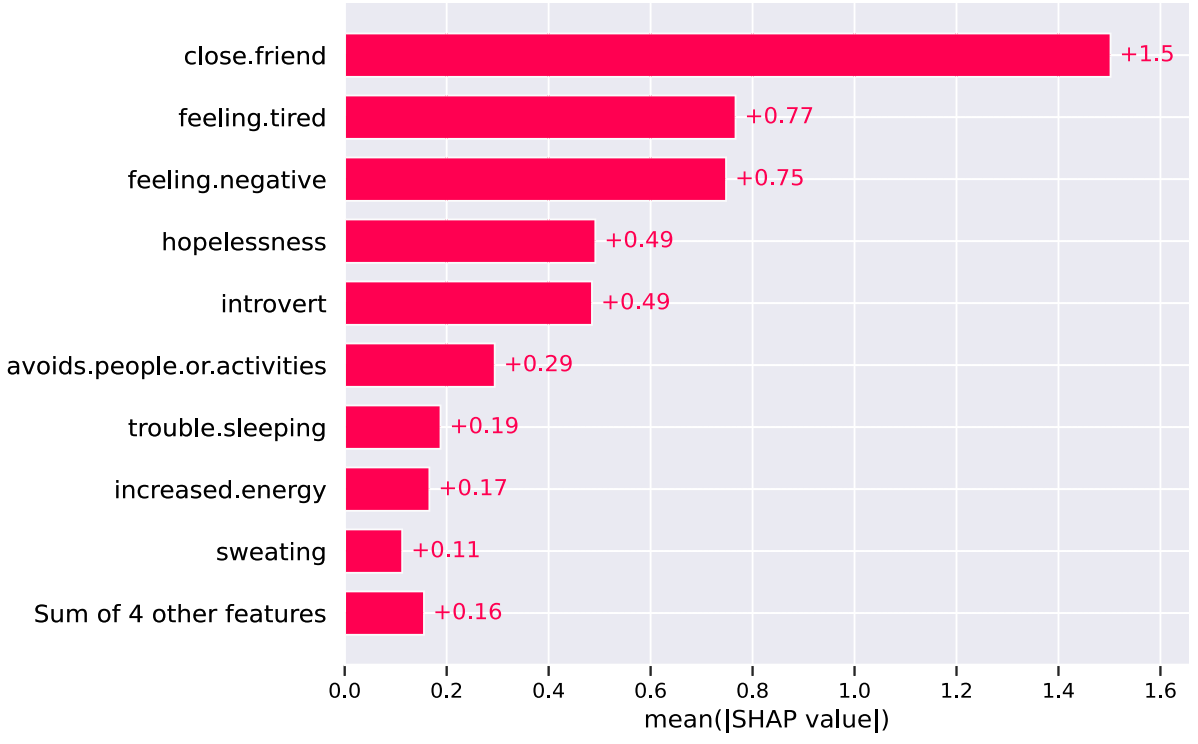

Fig. 9. The plots bar of the decision tree classifier.

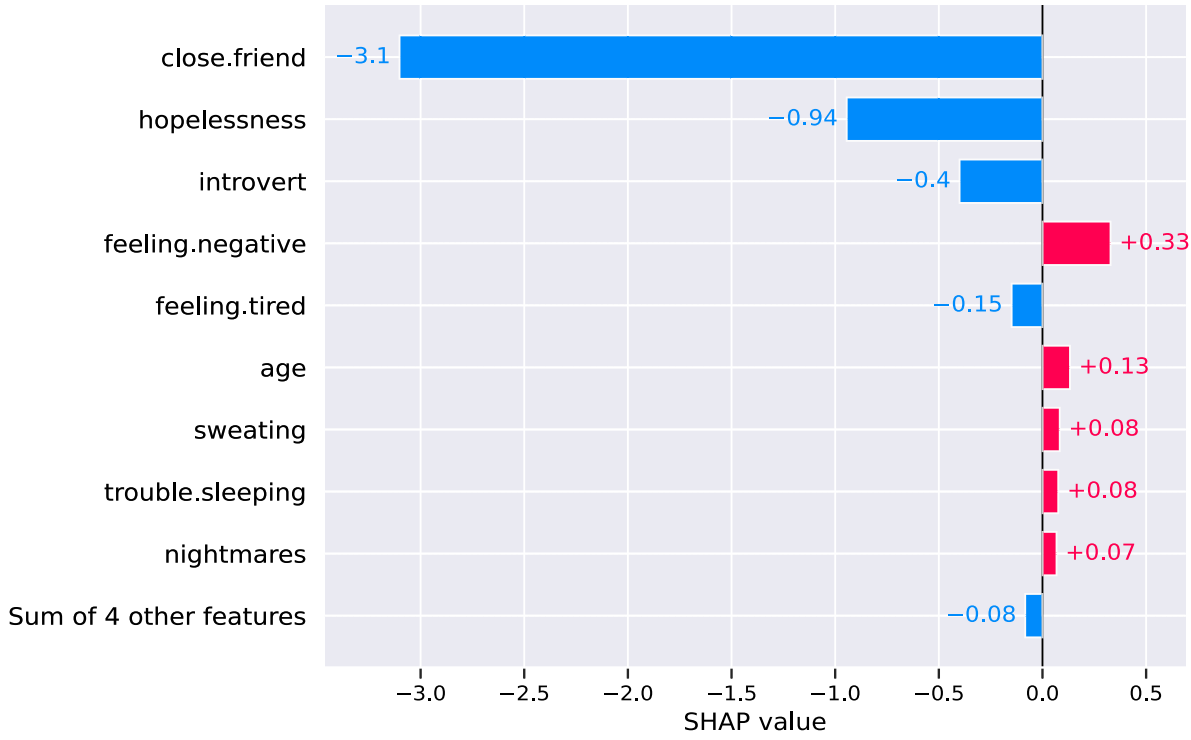

Fig. 10. The local bar plot bar of the decision tree classifier.

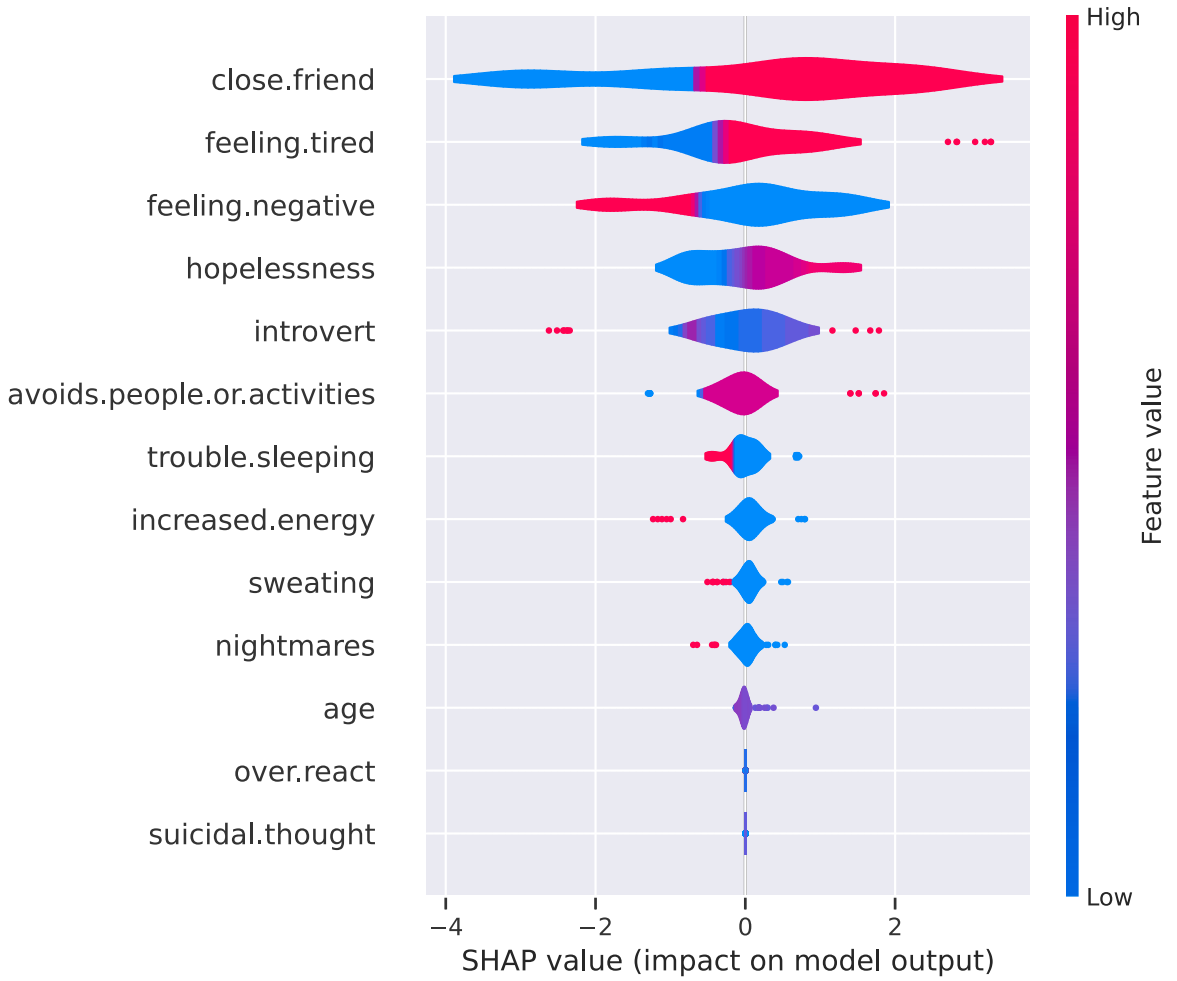

Fig. 11. The summary plot violin of the decision tree classifier.

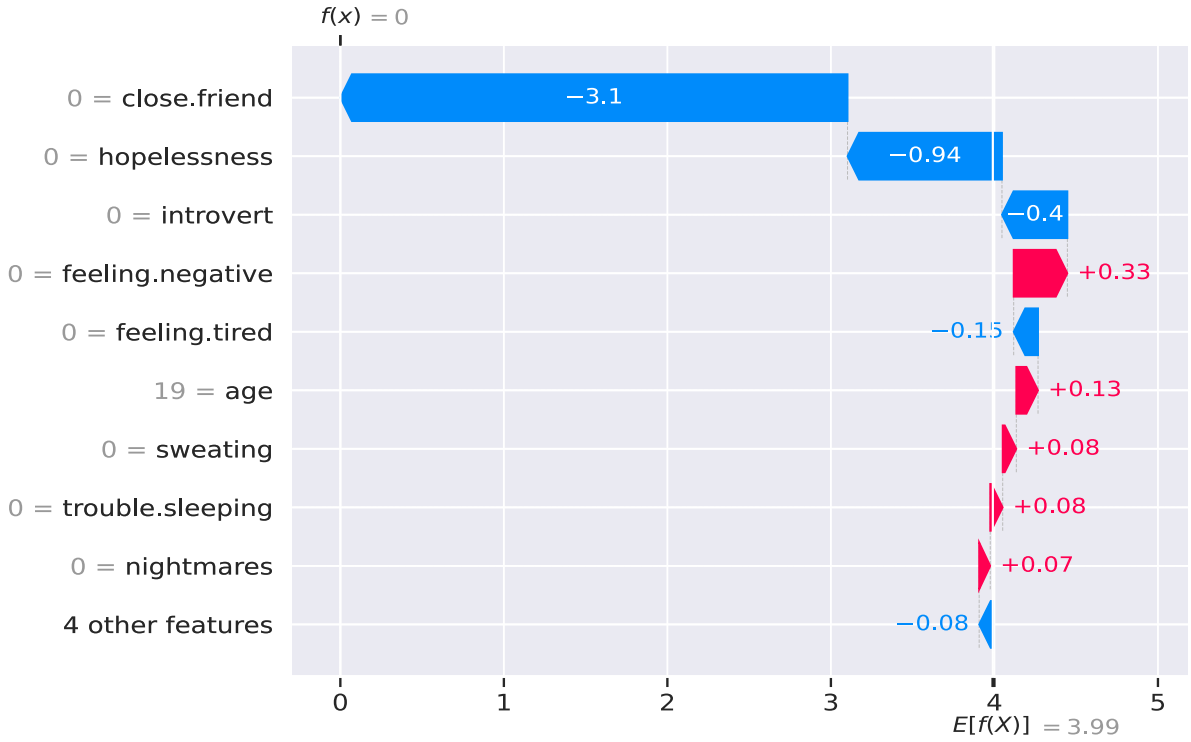

Fig. 12. The waterfall of decision tree classifier.

Table 4
Performance comparison of the proposed model against state-of-the-art models used for mental health disorders classification.

Model	Accuracy	Precision	Recall	F1-Score
KNN [13]	0.95	0.94	0.96	0.95
LSTM [13]	0.99	0.99	0.99	1.0
The proposed model: Voting(ANOVA, Mutinfo, RFE)	1.0	1.0	1.0	1.0

The performance of the proposed model is evaluated against state-of-the-art models for the classification of mental health disorders. Table 4 presents a comparison between the proposed stacking model and the KNN [13] and Long Short-Term Memory (LSTM) models [13]. The KNN model achieved an accuracy of 0.95, with precision, recall, and F1-score of 0.94, 0.96, and 0.95, respectively. The LSTM model demonstrated superior performance, achieving an accuracy of 0.99 and perfect scores (1.0) for precision, recall, and F1-score. The proposed stacking model, incorporating feature selection using ANOVA, mutual information, and recursive feature elimination, also achieved perfect scores across all metrics. According to the results, the stacking model outperforms the existing models.

The findings presented in this study contribute to the understanding of the classification model's performance and the influence of features on the predicted outcomes. By examining the confusion matrix, SHAP, and waterfall plot, key insights can be gained regarding the model's behavior and the significance of individual features. The confusion matrix of the SVC model provides a comprehensive evaluation of its performance across different classes. The accurate classification of instances along the diagonal reflects the model's ability to correctly assign instances to their respective classes. The SHAP technique offers a detailed understanding of the features' impact on the model's predictions. Analyzing the SHAP values allows for the identification of influential features that strongly contribute to the output. Positive and negative SHAP values indicate the direction of influence, providing valuable insights into feature relationships and their relative importance.

Using AI in mental healthcare requires careful consideration of ethical issues. Strong protections are needed to keep patient privacy and data safe, ensuring that data is only used for its original purpose. AI systems can sometimes be biased, unfairly affecting certain groups; ethical guidelines must address this problem. Therefore, AI tools should be simple to understand and have ways to fix issues so doctors and patients can trust them. Before utilizing the AI, patients (or their guardians) must have a comprehensive understanding of how it works and the potential risks. Finally, it is essential to make sure AI does not accidentally harm vulnerable people by, for example, creating inaccurate diagnoses or reinforcing harmful stereotypes.

Future research directions include enhancing model interpretability for broader clinical adoption, addressing ethical considerations, and validating the model's generalizability across diverse populations and clinical settings. Improving data quality and representativeness, particularly for autistic individuals across diverse demographics, is crucial. Furthermore, exploring hybrid models that combine different machine learning approaches (e.g., deep learning and rule-based systems) and incorporating additional data sources (e.g., physiological data, social media activity) could enhance predictive accuracy and interpretability. Finally, longitudinal studies are needed to assess the long-term predictive validity and stability of these models as individuals mature.

This work contributes to the growing field of AI-driven healthcare applications, particularly for neurodiverse populations. Finally, this study can address the pressing MHD crisis in Saudi Arabia and significantly improve early MHD diagnosis. The limitations of this study are that the sample size may not be large enough to fully represent the population, and the reliance on self-reported data could introduce biases. Additionally, the study focused on a specific set of features, and other relevant factors may not have been considered.

6. Conclusion

The recognition of mental disorder symptoms is crucial for timely management and reduction of recurring symptoms and disabilities. The ability to predict and explain mental health challenges can enable earlier intervention and more effective, individualized care plans,

improving the overall well-being of people with autism. This study presents a highly accurate and interpretable ensemble machine learning model for classifying mental health disorders, achieving perfect performance on all evaluation metrics. With an accuracy, precision, recall, and F1-Score of 1.0, this ensemble model demonstrates exceptional capabilities in accurately classifying mental health disorders. Furthermore, the model's explainable nature allows it to provide valuable insights into the obtained results. This feature enables clinicians and researchers to gain a deeper understanding of the factors contributing to disability in mental health disorders. Future research should focus on enhancing model generalizability, addressing ethical considerations, and improving data quality and diversity. The integration of diverse data sources and hybrid modeling approaches, coupled with longitudinal studies, will be crucial for establishing the long-term clinical utility and robustness of AI-driven mental health assessment and prediction.

CRediT authorship contribution statement

El-Sayed Atlam: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **M. Rokaya:** Writing – review & editing, Writing – original draft, Visualization, Resources, Project administration, Methodology, Data curation. **M. Masud:** Writing – review & editing, Writing – original draft, Validation, Software, Investigation, Formal analysis, Conceptualization. **H. Meshref:** Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Project administration, Formal analysis, Data curation. **Rakan Alotaibi:** Visualization, Validation, Software, Investigation, Formal analysis, Data curation. **Abdulqader M. Almars:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Data curation, Conceptualization. **Mohammed Assiri:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Funding acquisition, Data curation, Conceptualization. **Ibrahim Gad:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

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.

Acknowledgments

The authors extend their appreciation to the King Salman Center for Disability Research, Saudi Arabia for funding this work through Research Group no KSRG-2023-371, Funder id: http://dx.doi.org/10.13039/501100019345.

Data availability

https://www.kaggle.com/code/monicabackes/mental-disorders-data-analysis/input.

References

[1] M. Hamilton, Development of a rating scale for primary depressive illness, Br. J. Soc. Clin. Psychol. 6 (4) (1967) 278–296.
[2] WHO, Mental disorders, WHO, 2022, https://www.who.int/news-room/fact-sheets/detail/mental-disorders.
[3] V. Vitriol, A. Cancino, K. Weil, C. Salgado, M.A. Asenjo, S. Potthoff, et al., Depression and psychological trauma: An overview integrating current research and specific evidence of studies in the treatment of depression in public mental health services in Chile, Depression Res. Treat. 2014 (2014).
[4] A. Malki, E.-S. Atlam, A.E. Hassanien, A. Ewis, G. Dagnew, I. Gad, SARIMA model-based forecasting required number of COVID-19 vaccines globally and empirical analysis of peoples' view towards the vaccines, Alex. Eng. J. 61 (12) (2022) 12091–12110.
[5] K.S. Dobson, D.J. Dozois, Handbook of Cognitive-Behavioral Therapies, Guilford Publications, 2021.
[6] N.A. Baghdadi, S.M. Farghaly Abdelaliem, A. Malki, I. Gad, A. Ewis, E. Atlam, Advanced machine learning techniques for cardiovascular disease early detection and diagnosis, J. Big Data 10 (1) (2023) 144.
[7] C. Goodman, et al., American medical association council on scientific affairs, in: Medical Technology Assessment Directory: A Pilot Reference to Organizations, Assessments, and Information Resources, National Academies Press (US), 1988.
[8] I. Gad, M. Elmezain, M.M. Alwateer, M. Almaliki, G. Elmarhomy, E. Atlam, Breast cancer diagnosis using a machine learning model and swarm intelligence approach, in: 2023 1st International Conference on Advanced Innovations in Smart Cities, ICAISC, IEEE, 2023, pp. 1–5.
[9] R. Kakuma, H. Minas, N. Van Ginneken, M.R. Dal Poz, K. Desiraju, J.E. Morris, S. Saxena, R.M. Scheffler, Human resources for mental health care: current situation and strategies for action, Lancet 378 (9803) (2011) 1654–1663.
[10] A. Le Glaz, Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T.C. Ryan, J. Marsh, J. Devylder, M. Walter, S. Berrouiguet, et al., Machine learning and natural language processing in mental health: systematic review, J. Med. Internet Res. 23 (5) (2021) e15708.
[11] S. Graham, C. Depp, E.E. Lee, C. Nebeker, X. Tu, H.C. Kim, D.V. Jeste, Artificial intelligence for mental health and mental illnesses: an overview, Curr. Psychiatry Rep. 21 (2019) 1–18.
[12] A. Thieme, D. Belgrave, G. Doherty, Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems, ACM Trans. Comput.-Hum. Interact. 27 (5) (2020) 1–53.
[13] H. Alkahtani, T.H. Aldhyani, A.A. Alqarni, Artificial intelligence models to predict disability for mental health disorders, J. Disabil. Res. 3 (3) (2024) 20240022.
[14] T.H. Noor, A. Almars, I. Gad, E.S. Atlam, M. Elmezain, Spatial impressions monitoring during COVID-19 pandemic using machine learning techniques, Computers 11 (4) (2022) 52.
[15] S. Aleem, N.u. Huda, R. Amin, S. Khalid, S.S. Alshamrani, A. Alshehri, Machine learning algorithms for depression: diagnosis, insights, and research directions, Electronics 11 (7) (2022) 1111.
[16] M.K. Hooshmand, M.D. Huchaiah, A.R. Alzighaibi, H. Hashim, E.S. Atlam, I. Gad, Robust network anomaly detection using ensemble learning approach and explainable artificial intelligence (XAI), Alex. Eng. J. 94 (2024) 120–130.
[17] S. Hyman, D. Chisholm, R. Kessler, V. Patel, H. Whiteford, Mental disorders, in: Disease control priorities related to mental, neurological, developmental and substance abuse disorders, World Health Organization, 2006, pp. 1–20.
[18] A.A. Alhabeeb, R.A. Al-Duraihem, S. Alasmary, Z. Alkhamaali, N.A. Althumiri, N.F. BinDhim, National screening for anxiety and depression in Saudi Arabia 2022, Front. Public Heal. 11 (2023) 1213851.
[19] S.F. Jencks, Recognition of mental distress and diagnosis of mental disorder in primary care, Jama 253 (13) (1985) 1903–1907.
[20] C. Su, Z. Xu, J. Pathak, F. Wang, Deep learning in mental health outcome research: a scoping review, Transl. Psychiatry 10 (1) (2020) 116.
[21] N.K. Iyortsuun, S.-H. Kim, M. Jhon, H.J. Yang, S. Pant, A review of machine learning and deep learning approaches on mental health diagnosis, Healthcare 11 (3) (2023) 285.
[22] G. Arji, L. Erfannia, M. Hemmat, et al., A systematic literature review and analysis of deep learning algorithms in mental disorders, Informatics Med. Unlocked (2023) 101284.
[23] E. Garcia-Ceja, M. Riegler, T. Nordgreen, P. Jakobsen, K.J. Oedegaard, J. Tørresen, Mental health monitoring with multimodal sensing and machine learning: A survey, Pervasive Mob. Comput. 51 (2018) 1–26.
[24] T.H. Noor, A.M. Almars, A. El-Sayed, A. Noor, Deep learning model for predicting consumers' interests of IoT recommendation system, Int. J. Adv. Comput. Sci. Appl. 13 (10) (2022).
[25] S.E. Cho, Z.W. Geem, K.-S. Na, Prediction of depression among medical check-ups of 433,190 patients: A nationwide population-based study, Psychiatry Res. 293 (2020) 113474.
[26] A. Sau, I. Bhakta, Predicting anxiety and depression in elderly patients using machine learning technology, Heal. Technol. Lett. 4 (6) (2017) 238–243.
[27] J. Gratch, R. Artstein, G.M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, et al., The distress analysis interview corpus of human and computer interviews, in: LREC, Reykjavik, 2014, pp. 3123–3128.
[28] J. Yoon, C. Kang, S. Kim, J. Han, D-vlog: Multimodal vlog dataset for depression detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No.11, 2022, pp. 12226–12234.
[29] D. Xezonaki, G. Paraskevopoulos, A. Potamianos, S. Narayanan, Affective conditioning on hierarchical networks applied to depression detection from transcribed clinical interviews, 2020, arXiv preprint arXiv:2006.08336.
[30] A. Sharma, W.J. Verbeke, Improving diagnosis of depression with XGBOOST machine learning model and a large biomarkers Dutch dataset (n=11,081), Front. Big Data 3 (2020) 15.

[31] A.C. Timmons, J.B. Duong, N. Simo Fiallo, T. Lee, H.P.Q. Vo, M.W. Ahle, J.S. Comer, L.C. Brewer, S.L. Frazier, T. Chaspari, A call to action on assessing and mitigating bias in artificial intelligence applications for mental health, Perspect. Psychol. Sci. 18 (5) (2023) 1062–1096.

[32] E. Rocheteau, On the role of artificial intelligence in psychiatry, Br. J. Psychiatry 222 (2) (2023) 54–57.

[33] R. Abd Rahman, K. Omar, S.A.M. Noah, M.S.N.M. Danuri, M.A. Al-Garadi, Application of machine learning methods in mental health detection: a systematic review, IEEE Access 8 (2020) 183952–183964.

[34] E.-S. Atlam, M. Masud, M. Rokaya, H. Meshref, I. Gad, A.M. Almars, EASDM: Explainable autism spectrum disorder model based on deep learning, J. Disabil. Res. 3 (1) (2024) 20240003.

[35] M. Masud, A.M. Almars, M.B. Rokaya, H. Meshref, I. Gad, E.-S. Atlam, A novel light-weight convolutional neural network model to predict Alzheimer's disease applying weighted loss function, J. Disabil. Res. 3 (4) (2024) 20240042.

[36] A.M. Almars, Attention-based bi-LSTM model for arabic depression classification, Comput. Mater. Continua 71 (2) (2022).