ISSN 2442-6571

Soft voting technique with explainable artificial intelligence (XAI) for predicting depression, anxiety, and stress

Yefta Christian [a,1,*], Herman Herman [a,2], Muhammad Hafis [a,3], Kurnia Cantra [a,4]

[a] Universitas Internasional Batam, Indonesia
[1] yefta@uib.ac.id; [2] herman@uib.ac.id; [3] muhammad.hafis@uib.ac.id, [4] kurnia.cantra@uib.ac.id
* corresponding author

ARTICLE INFO

Article history
Received January 23, 2026
Revised February 26, 2026
Accepted March 16, 2026
Available online August 31, 2025

Keywords
Mental Health
DASS-42
Ensemble Learning
Soft Voting Classifier
Explainable AI

ABSTRACT

Mental health severity assessment is often hindered by limited access to professional services and the time required for clinical evaluation. This study proposes an interpretable soft voting method to classify the severity levels of depression, anxiety, and stress using DASS-42 questionnaire data. The proposed framework integrates Logistic Regression, Random Forest, Support Vector Machine, and Extreme Gradient Boosting, and is evaluated on 35,445 anonymized responses from a public psychometric dataset. Model performance was assessed using accuracy, precision, recall, F1-score, macro-averaged and weighted F1-score and the precision-recall curve under stratified cross-validation for class imbalance between the normal and mild classes. Explainable Artificial Intelligence using SHAP was employed to interpret model decisions. The soft voting achieved strong predictive performance, with accuracy values of 0.98 for depression, 0.99 for anxiety, and 0.97 for stress, outperforming or matching individual base models. SHAP analysis identified clinically consistent features contributing to model predictions; due to computational constraints, SHAP analysis was not applied to the SVM model. Despite strong performance, the use of secondary self-reported data and class imbalance, particularly the underrepresentation of normal and mild cases, were limitations. The proposed model demonstrates the potential of interpretable soft voting as a decision-support tool for mental health severity stratification in resource-constrained settings.

1. Introduction

Mental health conditions are a complex yet pervasive problem in the ever-changing world currently [1]. Understanding one's mental well-being is often overlooked and ignored due to social stigma and socio-cultural perception of said situation, even when the issue is affecting the day-to-day life of its sufferer [2]. Moreover, symptoms of mental health issues are hard to identify and predict due to the nature of the illness. Latest data released by the World Health Organization (WHO) stated that nearly one in seven people globally live with a mental disorder, with most individuals undertreated and health care for mental health often under-resourced [3]. Another meta-analysis done by [4] on the onset age of mental disorder across 192 countries found that the mental disorder onset peaks at 14.5 years old. In Indonesia specifically, mental health facilities and mental health practitioners are few and far between; sufferers are often perceived with negative stigma [5]. Additionally, high barriers to accessing mental health services, such as high cost, difficult access to facilities, delays in getting treatment, and a lack of professional facilitators, are some of the problems faced by Indonesian [6]. In a 2021 national survey, it's estimated that a third of all teenagers in Indonesia suffer from at least one mental health issue; anxiety disorder alone affects the majority of teenagers [7]. Additionally, the process of reaching a conclusive mental health assessment is time-consuming [8]. However, a quicker approach to get a glance at one's mental state is via the 42-Question Depression Anxiety Stress Scale (DASS-42), which is both reliable and consistent. DASS-42 is a clinically validated self-report psychometric tool designed to measure the severity of three related negative emotional states: depression, anxiety, and stress [9].

Artificial Intelligence (AI) technology in machine learning has advanced sufficiently and matured to play a significant role [10]. In particular, ensemble learning [11] has been used for postpartum depression predictive models [12], major depressive disorder (MDD) prediction [13], early detection of mental illness [14], predicting anxiety levels [15] and depression [16]. In its technicality, ensemble learning has been used in the form of decision trees [17], artificial neural networks [18], [19] to an adaptation of ensemble learning to analyze biometric primary data [20]. One of the types of ensemble learning is a voting classifier, which aggregates the predictions of several individual base models to make a final decision. One type of voting classifier is soft voting; the classifier averages the predicted probabilities assigned by each model to each class, allowing for a more nuanced final decision [21], which deals better with non-binary data. Despite the strong performance achieved by individual classifiers, namely SVM and Random Forest, several limitations remain, particularly in the presence of overlapping feature distributions and inconsistent detection of minority cases. Models trained independently tend to exhibit specific biases; SVM may enforce rigid decision boundaries, while Random Forest can be sensitive to data variability, resulting in misclassification when class characteristics overlap. These weaknesses indicate that reliance on a single learner may be insufficient for capturing the complex and nuanced patterns inherent in mental health data [22]. In particular, the soft voting classifier method offers a sophisticated way to refine these predictions by weighing the confidence levels of various underlying algorithms and combining the forecasts from multiple base models [23]. By aggregating the predictive strengths of diverse models, soft voting minimizes the bias or variance inherent in individual learners [24], resulting in a more resilient and clinically relevant decision-support tool that captures the complexities of human psychology better than traditional, monolithic approaches.

In the previous works, [25] uses the DASS-42 and min-max normalization to rescale and equalize the data into uniform range of 0-1 before analyzing the data with a set of ensemble classifier model using Convolutional Neural Network (CNN) to scan the features to classify the most relevant patterns and reducing noises, Recurrent Neural Network (RNN) for sequence modelling which fits perfectly for the DASS-42 data are a type of sequential data, and Long-Short Term-Memory to remember important information over very long information chain in which the DASS-42 sequences fits to this. The utilization of Large-scale Group Decision Making (LSGDM) acts to weight the prediction based on past performances and not only the consensus, and acts as the last optimizer and bias minimizer; the research also reached an accuracy of 98.88% and precision of 98.21%. However, the low interpretability of this research remains a gap.

Meanwhile, on top of DASS-42, demographic data were also used in this research by [26], who use a similar min-max normalization and ensemble voting classifier; however, they use Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) and integrate k-fold cross-validation (k=5), which increases robustness. This research reaches a mean accuracy of 94.5% with a 2.5% standard deviation. To validate, the paper uses the Patient-Health Questionnaire-9 (PHQ-9) alongside a Pearson correlation (with a coefficient of 0.82) with the model's predictions. Nevertheless, lower accuracy and demographic bias persisted in this research. Researchers in [27] also use DASS-42 with data pre-processing to standardize the data input by removing unnecessary data in the form of time taken to fill the questionnaire. The research uses machine learning techniques of SVM, RF, Decision Tree (DT), K-Nearest Neighbor (KNN), and Naive Bayes (NB) with k-fold cross-validation (k=5), which increases robustness. SVM achieved the highest accuracy of 99.3% for predicting depression, 98.9% for anxiety, and 98.8% for stress. However, the researcher particularly mentions that the SVM might suffer from a generalization gap between the training data and the learning data, which might lead to overfitting.

Nevertheless, issues persist with the utilization of a soft voting classifier in mental health. The lack of interpretability of soft voting models is notoriously hard to explain. Without understanding the logic, clinicians may find it difficult to trust the model's recommendation or explain the diagnosis to the patient. This issue is known as the "black-box" problem [28]. Additionally, dataset issues [29], [30] and the issue of reproducibility of the research [31], [32].

Advances in Artificial Intelligence (AI) and, to an extent, Explainable AI (XAI) have been the current state-of-the-art, with the usage of XAI in multiple fields for detection [33] and prediction [34] purposes. In particular, XAI has been used to increase transparency [35] in the field of healthcare, with the aim of creating a meaningful, impartial, and fair clinical diagnosis for mental healthcare practitioners. Building on these advancements, the integration of XAI is increasingly focused on bridging the gap that often hinders the clinical adoption of complex deep learning models.

Regardless, there exists a "black-box" situation, which is a complex and major issue when integrating AI with clinical systems [36]. The situation refers to the lack of transparency in how deep learning models arrive at specific conclusions. While an AI might accurately identify a mental health issue, the internal mathematical processes it uses are often so complex that even the developers who built the system cannot explain the exact logic behind a single output. This is where explainable AI (XAI) comes in as a form of interpretability to help practitioners "see" the underlying features and data points that drove the machine's decision. In a clinical setting, especially mental health, this is a matter of safety and professional ethics. By implementing XAI, we move from a "black-box" toward a "glass-box" model [37]. This transparency allows for a collaborative loop where the human expert can validate or override the AI's logic, ensuring that the final clinical decision is grounded in both data-driven insights and human intuition. Without this bridge, AI risks becoming a liability rather than a tool, potentially leading to misdiagnoses that could have been easily caught if the reasoning had been visible from the start. Additionally, this transparency allows easier interpretation for clinicians in their decision-making [38].

With the underlying problems mentioned, this research aims to use soft voting with base models of logistic regression, random forest, support vector machine, and XGBoost, alongside XAI, to improve the interpretability and accuracy of the proposed solution and mitigate the black-box problem. The aforementioned four base learners are chosen due to their accuracy performance on small-to-medium datasets [39]. Soft voting was chosen due to its capability and potency to improve accuracy compared to using only a single classifier machine learning technique [40]. With the mentioned research aim and contexts regarding the target demography, method employed, technique used, and clear reasoning, this research works towards answering these research questions as follows:

RQ1. To what extent does a soft-voting ensemble improve the prediction of DASS-42 severity levels by integrating the complementary decision boundaries learned by individual classifiers?

RQ2. How can explainable artificial intelligence (XAI) identify the most influential predictors of DASS-42 severity levels?

RQ3. How does the integration of the soft-voting ensemble with XAI enhance the interpretability of mental health assessments?

The research scope is limited to only the four base learning techniques of logistic regression, random forest, support vector machine, and XGBoost, along with soft voting. The research also limits the XAI interpretation to identification based on the DASS-42 for the machine learning training data.

2. Method

This study uses the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology [41] adapted to the research context to apply soft voting with base class models of logistic regression, random forest, support vector machine, and extreme gradient boosting, with XAI using SHAP to detect, interpret, and identify the results of the DASS-42 questionnaire.

2.1. System Architecture

Fig. 1 represents the system architecture of this research. The process begins with an understanding of healthcare practitioners to improve clinical outcomes and focuses on addressing the rising prevalence of depression, anxiety, and stress (DAS) in Indonesia by providing healthcare practitioners with a robust, interpretable decision-support tool. The primary objective is to bridge the gap between complex machine learning predictions and clinical utility. With that said, the data understanding starts with the raw DASS-42 data using a secondary source from Kaggle open psychometrics data, which also includes demographic factors and personality indicators.

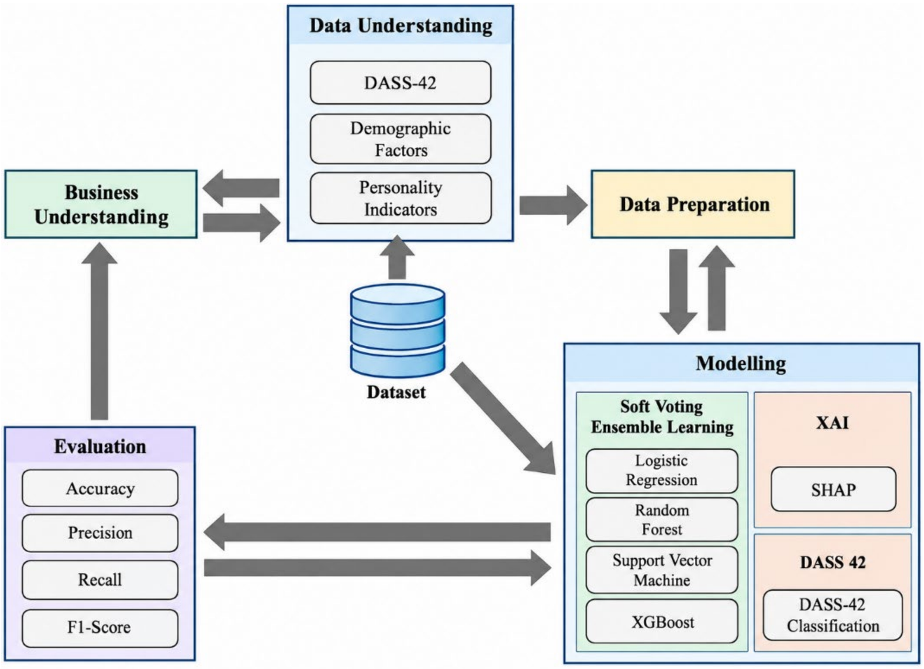


Fig. 1. Proposed system architecture using CRISP-DM methodology.

The data is then cleansed, preprocessed, and filtered to remove data anomalies. The data is then split into an 80:20 ratio, with 80% of the data used for training and 20% of the data used for testing the models used in this research. The models include Soft Voting, XAI using SHAP, and the classification of DASS-42, which are then evaluated. The evaluation phase assesses the performance of the soft voting with standard classification metrics, namely accuracy, precision, recall, F1-score, F1-score, F-1 score macro weight, and F1-score averaged. This phase then incorporates XAI via SHAP to interpret the model's decision-making process, allowing researchers and clinical practitioners to verify that the identified predictive features align with psychiatric domain knowledge.

2.2. Data Understanding

The datasets used in the study were public datasets available in Depression Anxiety Stress Scales Responses from Kaggle, which were hosted on OpenPsychometrics.org, with a total of 39,775 self-report survey responses, consisting of a combination of questions to assess depression, anxiety, and stress, represented in Q1A-Q42A. Additionally, these datasets included personality traits and demographic information regarding respondents' personal information, namely education, country, age, gender, English native and urban; most respondents are young, aged 20-30 and gender of female. As the name suggests, DASS-42 utilizes 42 questions and a scale of 0-3, weighting each question to a corresponding variable.

2.3. Data Preparation

During the data preparation phase, the data is pre-processed to remove outliers and missing values. The dataset is filtered with a starting dataset of 39,775 down to 35,445. This process is done to ensure the quality of the dataset for DASS-42 classification of depression, anxiety, and stress, with each having

14 questions. Following the data cleansing process, data wrangling was performed to improve model interpretability and preprocessing efficiency. This included renaming DASS-related columns to enhance SHAP interpretability, as well as encoding categorical variables alongside their category interpretation of Depression/Anxiety/Stress, as shown in Table 1.

Table 1. DASS-42 based on [42] to SHAP Interpretability Table.

Q No.	Question	SHAP Interpretability	D/A/S
Q1A	I found myself getting upset by quite trivial things	Trivial Upset	S
Q2A	I was aware of the dryness of my mouth	Dry Mouth	A
Q3A	I couldn't seem to experience any positive feeling at all	No Positivity	D
Q4A	I experienced breathing difficulty (eg, excessively rapid breathing, breathlessness in the absence of physical exertion)	Breathing Difficulty	A
Q5A	I just couldn't seem to get going	Lack of Drive	D
Q6A	I tended to overreact to situations	Overreaction	S
Q7A	I had a feeling of shakiness (eg, legs going to give way)	Shakiness	A
Q8A	I found it difficult to relax	Difficulty Relaxing	S
Q9A	I found myself in situations that made me so anxious I was most relieved when they ended	Situational Anxiety	A
Q10A	I felt that I had nothing to look forward to	Hopelessness	D
Q11A	I found myself getting upset rather easily	Emotional Sensitivity	S
Q12A	I felt that I was using a lot of nervous energy	Nervous Energy	S
Q13A	I felt sad and depressed	Depressed Mood	D
Q14A	I found myself getting impatient when I was delayed in any way (eg, lifts, traffic lights, being kept waiting)	Impatience	S
Q15A	I had a feeling of faintness	Faintness	A
Q16A	I felt that I had lost interest in just about everything	Loss of Interest	D
Q17A	I felt I wasn't worth much as a person	Low Self-Worth	D
Q18A	I felt that I was rather touchy	Touchiness	S
Q19A	I perspired noticeably (eg, hands sweaty) in the absence of high temperatures or physical exertion	Excessive Sweating	A
Q20A	I felt scared without any good reason	Unexplained Fear	A
Q21A	I felt that life wasn't worthwhile	Life Worthlessness	D
Q22A	I found it hard to wind down	Trouble Winding Down	S
Q23A	I had difficulty in swallowing	Swallowing Difficulty	A
Q24A	I couldn't seem to get any enjoyment out of the things I did	Anhedonia	D
Q25A	I was aware of the action of my heart in the absence of physical exertion (eg, sense of heart rate increase, heart missing a beat)	Heart Awareness	A
Q26A	I felt down-hearted and blue	Low Mood	D
Q27A	I found that I was very irritable	Irritability	S
Q28A	I felt I was close to panic	Panic Feelings	A
Q29A	I found it hard to calm down after something upset me	Difficulty Calming	S
Q30A	I feared that I would be 'thrown' by some trivial but unfamiliar task	Task Anxiety	A
Q31A	I was unable to become enthusiastic about anything	Lack of Enthusiasm	D
Q32A	I found it difficult to tolerate interruptions to what I was doing	Interruption Intolerance	S
Q33A	I was in a state of nervous tension	Nervous Tension	S
Q34A	I felt I was pretty worthless	Worthlessness	D
Q35A	I was intolerant of anything that kept me from getting on with what I was doing	Frustration Intolerance	S
Q36A	I felt terrified	Terror	A
Q37A	I could see nothing in the future to be hopeful about	Future Hopelessness	D
Q38A	I felt that life was meaningless	Life Meaninglessness	D
Q39A	I found myself getting agitated	Agitation	S
Q40A	I was worried about situations in which I might panic and make a fool of myself	Panic Worry	A
Q41A	I experienced trembling (eg, in the hands)	Trembling	A
Q42A	I found it difficult to work up the initiative to do things	Low Initiative	D

Afterwards, the country variable, which contained 134 unique categories, was converted from categorical text to numerical values. To mitigate the issue of dimensionality, LabelEncoder was applied instead of One-Hot Encoding, as the latter could negatively impact model training [43].

Upon completion of the preprocessing stage, feature selection, and dimensionality reduction. The selected features, namely personality traits and demographic information regarding respondents' personal information, namely country, age, gender, and English native, urban, corresponded to the three DASS-42 constructs: depression, anxiety, and stress, were trained individually with an 80/20 train-test data split. The 80/20 split ensures training quality [44] and accurate prediction on five levels of severity, namely normal, mild, moderate, severe and extremely severe, based on [42], with StratifiedKFold employed to mitigate class imbalance.

2.4. Modelling

The voting classifier combines predictions from multiple base classifiers to generate a final prediction, thereby leveraging the strength of the individual models [40]; specifically, four classifiers were used as base learners: Logistic Regression, Random Forest, Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost). Soft voting is applied to determine the final class label probabilities from each base classifier. Logistic Regression model utilizes the sigmoid function as a probabilistic linear model that establishes a relationship between features and the target classes, with a hyperparameter of C = 10 to control regularization strength and prevent overfitting DASS-42 dataset, a higher value of C tells the model to give more weight to the training data while maintaining an L2 penalty.

Random Forest model takes advantage of multitude of decision trees to create probability of all trees, with tuning of n_estimators = 100, max_depth = 10, and min_sample_split = 2, n_estimators are the number of tree used in random forest model, with max_depth of 10, which limit model depth to reduce overfitting, and min_sample_split dictate that model need to have at least two sample to split, constrain were use to ensure that model captures significant pattern without memorizing. Support Vector Machine algorithms process information by mapping it into high-dimensional features to identify an optimal hyperplane in training, using C as regularization; with C = 10, regularization of misclassification increases, meaning that the model has low tolerance for mistakes, lowering bias; it increases variance. XGBoost model predicts the target value using an ensemble of weak learners to minimize the loss function, tuning n_estimators = 100 and max_depth = 10, with only 100 trees used and limiting depth to reduce overfitting and maintain L1 and L2 regularization and to reduce overall model overfitting and bias issues; implementation of Stratified K-Fold with 5-fold cross-validation to compensate for the imbalanced dataset [45]. The details of the hyperparameter tuning following previous literature [26] can be viewed in Table 2.

Table 2. Model Hyperparameter Tuning.

Individual Model	Hyperparameter Tuning on DASS-42
Logistic Regression	C = 10
Random Forest	n_estimators = 100, max depth = 10, min_sample_split = 2
Support Vector Machine	C = 10
XGBoost	n_estimators = 100, max depth = 10

Regarding the interpretation of XAI, SHapley Additive exPlanations (SHAP) [46] is used. SHAP is a mathematical framework used in XAI to determine exactly how much each input variable (feature) contributes to a specific prediction made by a machine learning model. SHAP calculates the contribution of each feature by considering all possible combinations of features, which gives SHAP the capability to provide global and local explanations, as well as explain non-linear relationships [47]. In practical terms, this means SHAP helps us look inside the black box model to see why it made a specific choice. By assigning a specific numeric value to each feature's influence, it transforms abstract algorithmic logic into a prioritized list of drivers that humans can interpret.

3. Results and Discussion

3.1. Evaluation

To further evaluate the model's classification prediction and accuracy, researchers use classification metrics, namely accuracy, precision, recall, F1-score, F1-score macro, F1-score weighted and support. Multiple metric evaluation ensures that model performance is not solely based on a single variable (Table 3). The Soft Voting approach achieved consistently high performance across all three conditions, with scores of 0.98 for depression, 0.99 for anxiety, and 0.97 for stress. For depression, the Support Vector Machine achieved the highest individual score of 0.99, slightly outperforming Soft Voting. For anxiety, Soft Voting matched the best-performing individual model, Extreme Gradient Boosting, with a score of 0.99. In stress classification, the Support Vector Machine obtained the highest score of 0.98, followed closely by Soft Voting at 0.97. These results indicate that although Soft Voting does not outperform the best individual classifier in every category, it provides strong and consistent performance across all three mental health conditions, highlighting the potential benefit of combining multiple classifiers in an ensemble framework..

Table 3. Base Class Models vs Soft Voting.

Mental Health Condition	Logistic Regression	Random Forest	Support Vector Machine	Extreme Gradient Boosting	Soft Voting
Depression	0.92	0.96	0.99	0.98	**0.98**
Anxiety	0.93	0.98	0.97	0.99	**0.99**
Stress	0.83	0.92	0.98	0.96	**0.97**

The macro-averaged F1-score: Logistic Regression and Random Forest achieved the highest performance, both with a score of 0.98, while Support Vector Machine obtained a slightly lower score of 0.96 (Table 4). For the weighted-average F1-score, Random Forest achieved the highest score of 0.99, followed by Logistic Regression at 0.98 and Support Vector Machine at 0.97. The consistently high macro and weighted F1-scores indicate that all three classifiers provide strong classification performance. The small differences between the macro and weighted averages also suggest relatively stable performance across classes, with Random Forest showing the strongest overall results among the evaluated classifiers

Table 4. Voting Classifier Macro and Weighted F1-Score.

Evaluation Metric	Logistic Regression	Random Forest	Support Vector Machine
F1-Score Macro Averaged	0.98	0.98	0.96
F1-Score Weighted Average	0.98	0.99	0.97

As shown in Table 5, most respondents have higher depression; an imbalance of data occurs in prediction, with no respondents having normal or mild depression, mostly moderate, severe, and extremely severe, which hinders the prediction property of the voting classifier, as the model found no data to train upon. Fig. 2 shows the depression Precision-Recall curve and confusion matrix, as the confusion matrix shows LabelEncoder alphabetical order, namely Extremely Severe, Mild, Moderate, Normal and Severe, excluding AP, as AP follows the Table 5 format of Moderate, Severe, and Extremely Severe, the consequence of low AP and misclassification in the Severe class, with metrics.

Table 5. Depression (Class Base).

	0 (Normal)	1 (Mild)	2 (Moderate)	3 (Severe)	4 (Extremely Severe)
Precision			0.98	0.96	0.99
Recall			0.97	0.94	0.99
F1-Score			0.98	0.96	0.99
Support			5455	5881	24109

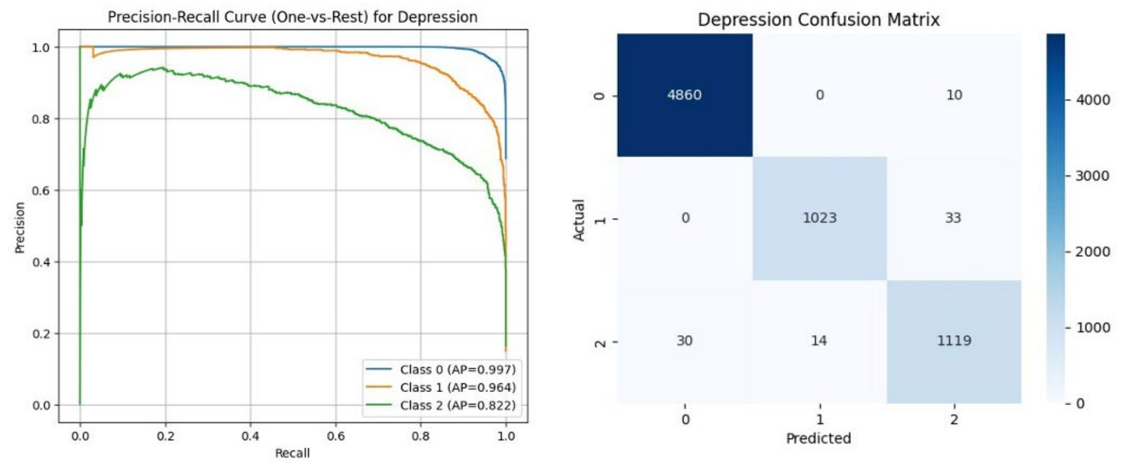


Fig. 2. Precision-Recall Curve and Confusion Matrix on Depression.

Imbalance also occurs in the anxiety dataset, shown in Table 6; more respondents are prone to anxiety, with many respondents suffering from extremely severe anxiety, distributed among severe and followed by moderate anxiety. Fig. 3 shows the AP and anxiety confusion matrix; the configuration of the AP and confusion matrix is identical to that of depression, with the latter Severe class being misclassified and low AP corresponding to the Moderate class metrics.

Table 6. Anxiety (Class Base).

	0 (Normal)	1 (Mild)	2 (Moderate)	3 (Severe)	4 (Extremely Severe)
Precision			0.98	0.97	0.99
Recall			0.96	0.97	0.99
F1-Score			0.97	0.97	0.99
Support			808	5364	29273

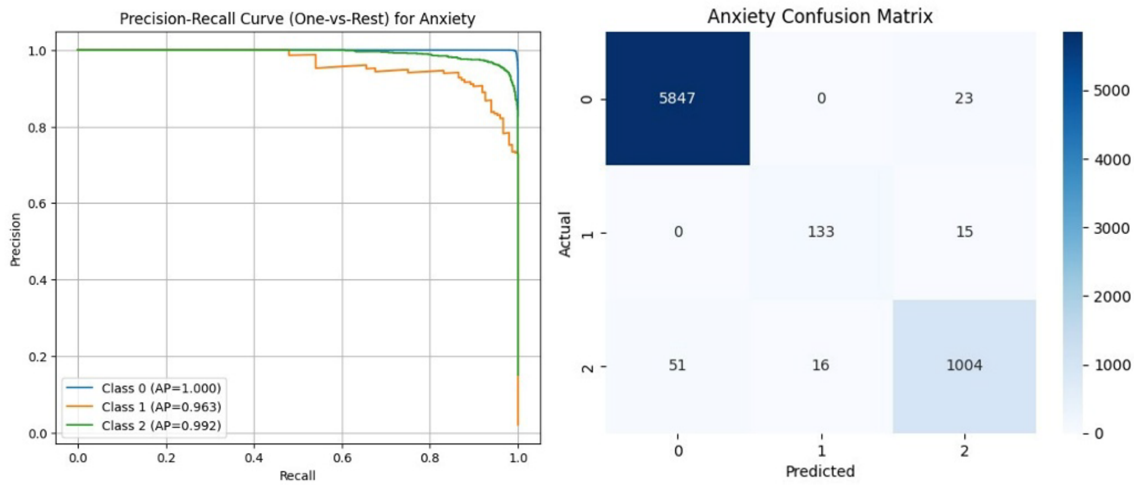


Fig. 3. Precision-Recall Curve and Confusion Matrix for Anxiety.

The distribution of data in stress is much more distributed in comparison with depression and anxiety, as shown in Table 7, in which models are better at predicting stress; with k-folding, metric output is more balanced, even when datasets are imbalanced, which increases model robustness in prediction. Fig. 4 shows AP and the confusion matrix with the configuration of Extremely Severe, Mild, Moderate, Normal and Severe, and AP follows Table 7, with less misclassification and high AP, as data are distributed in each class.

Table 7. Stress (Class Base).

	0 (Normal)	1 (Mild)	2 (Moderate)	3 (Severe)	4 (Extremely Severe)
Precision	1	0.96	0.95	0.97	0.99
Recall	0.90	0.90	0.96	0.97	0.99
F1-Score	0.95	0.93	0.95	0.97	0.99
Support	263	1727	5500	8515	19440

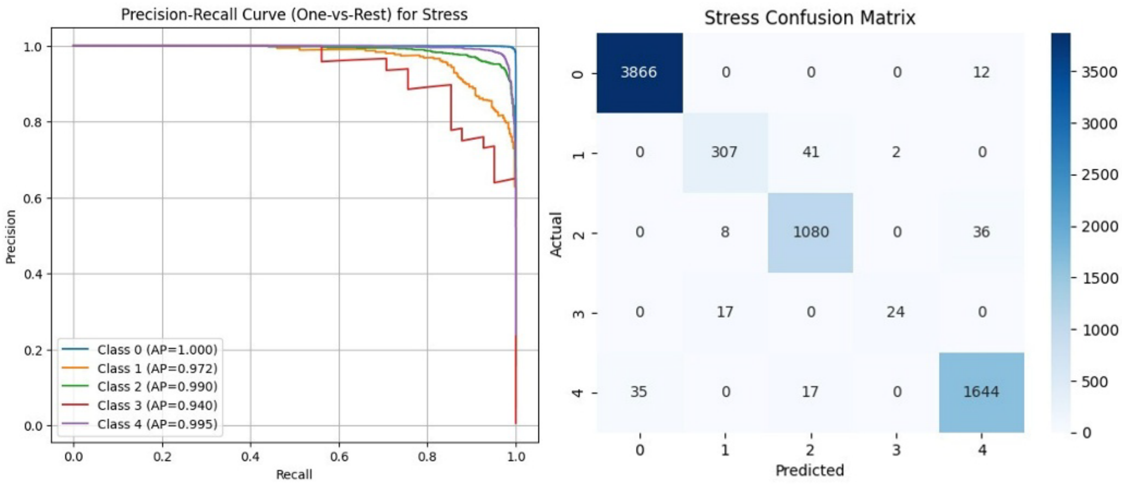


Fig. 4. Precision-Recall Curve and Confusion Matrix on Stress.

The lack of "Normal" or "Mild" cases in the dataset may reflect a sampling bias towards individuals already seeking mental health support or the high prevalence of psychological distress within the demography.

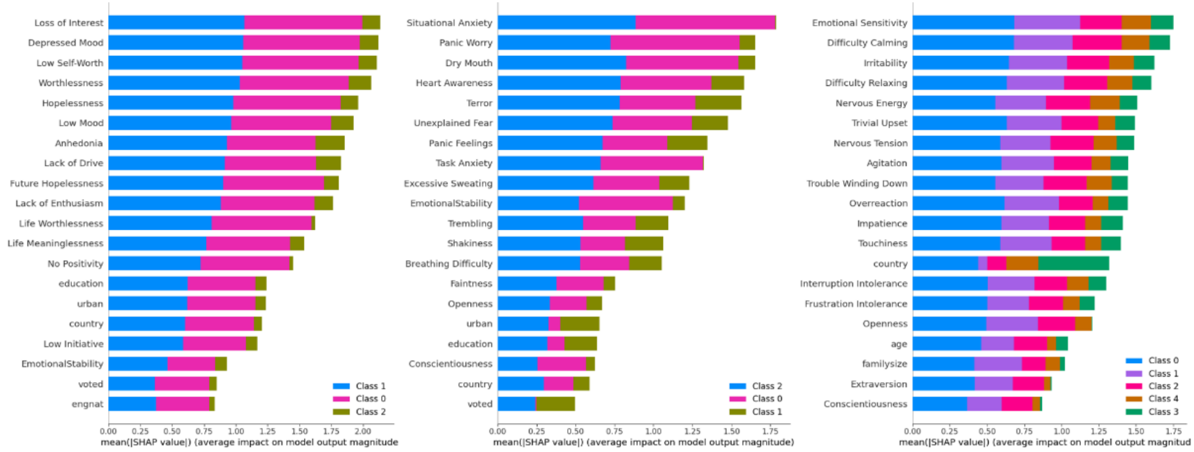
3.2. XAI Interpretation

SHAP provides both global and local interpretability. Global explanation identifies overall feature importance across the dataset, revealing patterns learned by the model, while local explanation explains individual predictions by quantifying the contribution of each feature for a specific instance. This dual interpretability enhances model transparency and supports its use as a reliable decision-support tool.

The evaluation metrics presented in Tables 3 to 7 have provided quantitative evidence of the models' performance and reliability. These metrics, namely accuracy, precision, recall, and F1-score, indicate how well the model performs. On the contrary, there is no information on how the models create predictions. In other words, while scores reflect the outcome of the learning process, they do not reveal the internal decision-making mechanisms of the model.

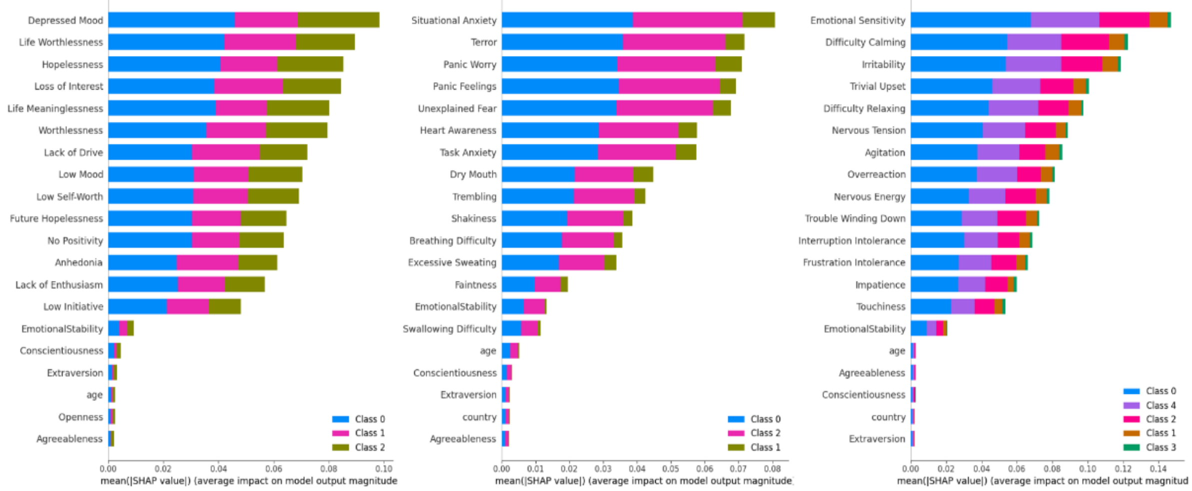
While SVM achieved the highest overall performance across evaluation metrics, SHAP results for this model could not be presented, as KernelExplainer was unable to reliably compute SHAP values for the full dataset. This limitation arises from the poor scalability of KernelExplainer with large sample sizes, rendering the computation of SHAP values for all 35,445 instances infeasible within a reasonable time frame. In practice, the explanation process exhibited excessive computational overhead, with runtimes extending to several days, frequent memory exhaustion leading to kernel crashes, and increasingly unstable or noisy SHAP estimations due to the reliance on extensive sampling. Consequently, the application of SHAP to large-scale SVM models proved impractical, despite the model's superior predictive performance.

In Figs. 5 to 8, SHAP values for classes 0, 1 and 2 (Figures a and b) correspond to classes 2 (Moderate), 3 (Severe) and 4 (Extremely Severe) from Table 2 and Table 3, this is because class 0 (Normal) and class 1 (Mild) are absent from the dataset. Although class imbalance is observed, the dataset shows no class imbalance. While each model learns distinct parameter configurations, they exhibit substantial common ground in terms of influential features. As the Logistic Regression, Random Forest and XGBoost models learned, the SHAP labels found a commonality in influential features used for model decision-making; these were Depressed Mood, Low Self-Worth and Hopelessness in the Depression model.



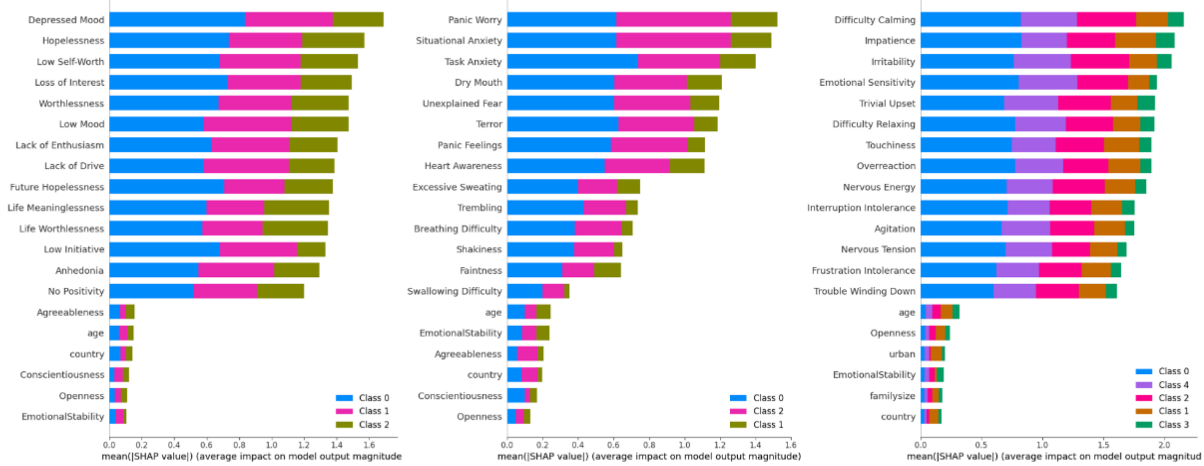
(a) (b) (c)

Fig. 5. Logistic Regression SHAP Interpretation of a) Depression, b) Anxiety, and c) Stress.



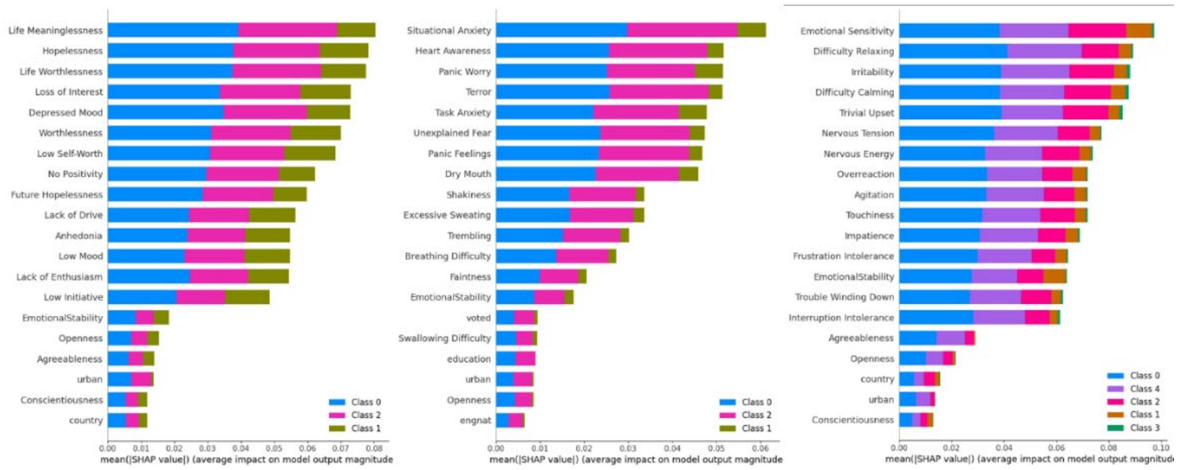
(a) (b) (c)

Fig. 6. Random Forest SHAP Interpretation of a) Depression, b) Anxiety, and c) Stress.



(a) (b) (c)

Fig. 7. XGBoost SHAP Interpretation of a) Depression, b) Anxiety, and c) Stress.



(a) (b) (c)

Fig. 8. Soft Voting SHAP Interpretation of a) Depression, b) Anxiety, and c) Stress.

Table 8 presents the XAI-based feature interpretations across the classification models for depression, anxiety, and stress. The results reveal consistent patterns across models, with depressive mood and hopelessness emerging as key features for depression, situational anxiety and panic-related worry for anxiety, and difficulty calming and irritability for stress. These findings indicate that, despite differences in model architecture, the models generally rely on similar symptom-related features when making predictions.

Table 8. XAI Feature Interpretation.

Model	DASS	Feature
Logistic Regression	Depression	Loss of Interest,Depressed Mood, Low Self-Worth
Logistic Regression	Anxiety	Situational Anxiety, Panic Worry, Dry Mouth
Logistic Regression	Stress	Emotional Sensitivity, Difficulty Calming, Irritability
Random Forest	Depression	Depressed Mood, Life Worthlessness, Hopelessness
Random Forest	Anxiety	Situational Anxiety, Terror, Panic Worry
Random Forest	Stress	Emotional Sensitivity, Difficulty Calming, Irritability
Extreme Gradient Boosting	Depression	Depressed Mood, Hopelessness, Low Self-Worth
Extreme Gradient Boosting	Anxiety	Panic Worry, Situational Anxiety, Task Anxiety
Extreme Gradient Boosting	Stress	Difficulty Calming, Impatience, Irritability
Soft Voting	Depression	Life Meaninglessness, Hopelessness, Life Worthlessness
Soft Voting	Anxiety	Situational Anxiety, Heart Awareness, Panic Worry
Soft Voting	Stress	Emotional Sensitivity, Difficulty Relaxing, Irritability

4. Conclusion

This study developed an interpretable soft voting framework combining Logistic Regression, Random Forest, Support Vector Machine, and XGBoost to classify DASS-42 severity levels of depression, anxiety, and stress. Using the CRISP-DM methodology and Stratified 5-fold cross-validation, the ensemble achieved high accuracy, namely 0.98 for depression, 0.99 for anxiety, and 0.97

for stress and demonstrated more stable performance than individual models under class imbalance. However, performance should be interpreted cautiously due to the strong skew toward moderate-to-extremely severe cases and the absence of normal and mild classes in parts of the dataset. SHAP-based analysis showed consistent feature patterns across models. Individual classifiers mainly emphasized direct symptom-level indicators such as depressed mood, panic worry, and difficulty calming. In contrast, the soft voting ensemble highlighted broader cognitive and existential constructs, including life meaninglessness, hopelessness, and life worthlessness. This suggests that aggregating complementary decision boundaries can produce more clinically meaningful explanations. Notably, although SVM achieved the highest predictive performance, SHAP explanations for SVM were computationally infeasible at full scale, representing a methodological limitation. Several limitations must be acknowledged. The study relied on secondary self-reported data from an online source, introducing potential reporting and sampling bias. Significant class imbalance, particularly the underrepresentation or absence of normal and mild cases, restricts generalizability. Furthermore, predictions were not externally validated against independent clinical diagnoses. Therefore, the proposed framework should be regarded as a decision-support tool for screening and stratification rather than a diagnostic system. Future research should incorporate balanced and clinically diverse datasets, perform external clinical validation and calibration analysis, and explore ordinal classification approaches aligned with severity structure. Addressing computational constraints in explaining high-performing models such as SVM would further strengthen methodological rigor and clinical applicability.

Acknowledgment

We would like to express our gratitude to the Faculty of Computer Science and the Department of Information Systems at Universitas International Batam for their support of this study.

Declarations

Author contribution. Conceptualization: Yefta Christian, Kurnia Cantra; Methodology: Yefta Christian, Kurnia Cantra; Writing - original draft: Yefta Christian, Kurnia Cantra, Muhammad Hafis; Writing-review & editing: Herman.

Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper

References

[1] V. Singh, A. Kumar, and S. Gupta, "Mental Health Prevention and Promotion—A Narrative Review," *Front. Psychiatry*, vol. 13, no. July, p. 898009, Jul. 2022, doi: 10.3389/fpsyt.2022.898009.

[2] K. K. Y. Ma, J. K. Anderson, and A. Burn, "Review: School-based interventions to improve mental health literacy and reduce mental health stigma – a systematic review," *Child Adolesc. Ment. Health*, vol. 28, no. 2, pp. 230–240, May 2023, doi: 10.1111/camh.12543.

[3] "World health statistics 2025: monitoring health for the SDGs, Sustainable Development Goals," World Health Organization, 2025. [Online]. Available at: https://www.who.int/publications/b/78420.

[4] M. Solmi *et al.*, "Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies," *Mol. Psychiatry*, vol. 27, no. 1, pp. 281–295, Jan. 2022, doi: 10.1038/s41380-021-01161-7.

[5] M. A. Subu *et al.*, "Exploring mental health stigma among Indonesian healthcare students towards individuals with mental illnesses: a qualitative study," *Int. J. Qual. Stud. Health Well-being*, vol. 19, no. 1, p. 11, Dec. 2024, doi: 10.1080/17482631.2024.2327103.

[6] L. Munira, P. Liamputtong, and P. Viwattanakulvanid, "Barriers and facilitators to access mental health services among people with mental disorders in Indonesia: A qualitative study," *Belitung Nurs. J.*, vol. 9, no. 2, pp. 110–117, Apr. 2023, doi: 10.33546/bnj.2521.

[7] A. E. Wahdi *et al.*, "Indonesia – National Adolescent Mental Health Survey (I-NAMHS) Report (Bahasa Indonesia) - Queensland Centre for Mental Health Research (QCMHR)," QCMHR-Queensland Centre

for Mental Health Research. Accessed: Mar. 31, 2026. [Online]. Available at: https://qcmhr.org/outputs/reports/12-i-namhs-report-bahasa-indonesia.

[8] M. Subotic-Kerry *et al.*, "While they wait: a cross-sectional survey on wait times for mental health treatment for anxiety and depression for adolescents in Australia," *BMJ Open*, vol. 15, no. 3, p. e087342, Mar. 2025, doi: 10.1136/bmjopen-2024-087342.

[9] M. Makara-Studzińska, E. Tyburski, M. Załuski, K. Adamczyk, J. Mesterhazy, and A. Mesterhazy, "Confirmatory Factor Analysis of Three Versions of the Depression Anxiety Stress Scale (DASS-42, DASS-21, and DASS-12) in Polish Adults," *Front. Psychiatry*, vol. 12, p. 770532, Jan. 2022, doi: 10.3389/fpsyt.2021.770532.

[10] T. Timakum, Q. Xie, and M. Song, "Analysis of E-mental health research: mapping the relationship between information technology and mental healthcare," *BMC Psychiatry*, vol. 22, no. 1, p. 57, Dec. 2022, doi: 10.1186/s12888-022-03713-9.

[11] R. Islam and M. A. Layek, "StackEnsembleMind: Enhancing well-being through accurate identification of human mental states using stack-based ensemble machine learning," *Informatics Med. Unlocked*, vol. 43, no. 2, p. 101405, Jan. 2023, doi: 10.1016/j.imu.2023.101405.

[12] W. A. Fazraningtyas, B. Rahmatullah, H. Naparin, M. Basit, and N. A. Razak, "A predictive model for postpartum depression: ensemble learning strategies in machine learning," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 37, no. 1, p. 443, Jan. 2025, doi: 10.11591/ijeecs.v37.i1.pp443-451.

[13] Z. Chen *et al.*, "Identifying major depressive disorder among US adults living alone using stacked ensemble machine learning algorithms," *Front. Public Heal.*, vol. 13, p. 1472050, Feb. 2025, doi: 10.3389/fpubh.2025.1472050.

[14] K. Vaishnavi, U. Nikhitha Kamath, B. Ashwath Rao, and N. V Subba Reddy, "Predicting Mental Health Illness using Machine Learning Algorithms," *J. Phys. Conf. Ser.*, vol. 2161, no. 1, p. 012021, Jan. 2022, doi: 10.1088/1742-6596/2161/1/012021.

[15] A. Daza, Arroyo-Paz, J. Bobadilla, O. Apaza, and J. Pinto, "Stacking ensemble learning model for predict anxiety level in university students using balancing methods," *Informatics Med. Unlocked*, vol. 42, no. 4, p. 101340, Jan. 2023, doi: 10.1016/j.imu.2023.101340.

[16] A. Daza Vergaray, J. C. H. Miranda, J. B. Cornelio, A. R. López Carranza, and C. F. Ponce Sánchez, "Predicting the depression in university students using stacking ensemble techniques over oversampling method," *Informatics Med. Unlocked*, vol. 41, p. 101295, Jan. 2023, doi: 10.1016/j.imu.2023.101295.

[17] K. Battista, L. Diao, K. A. Patte, J. A. Dubin, and S. T. Leatherdale, "Examining the use of decision trees in population health surveillance research: an application to youth mental health survey data in the COMPASS study," *Heal. Promot. Chronic Dis. Prev. Canada*, vol. 43, no. 2, pp. 73–86, Feb. 2023, doi: 10.24095/hpcdp.43.2.03.

[18] S. A. Sulak and N. Koklu, "Analysis of Depression, Anxiety, Stress Scale (DASS-42) With Methods of Data Mining," *Eur. J. Educ.*, vol. 59, no. 4, p. e12778, Dec. 2024, doi: 10.1111/ejed.12778.

[19] A. Siddiqua, A. M. Oni, A. Saleh Musa Miah, and J. Shin, "Enhancing PTSD Outcome Prediction with Ensemble Models in Disaster Contexts," in *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, IEEE, Feb. 2025, pp. 1–6. doi: 10.1109/ECCE64574.2025.11013125.

[20] S. Hadhri, M. Hadiji, and W. Labidi, "A voting ensemble classifier for stress detection," *J. Inf. Telecommun.*, vol. 8, no. 3, pp. 399–416, Jul. 2024, doi: 10.1080/24751839.2024.2306786.

[21] K. Akyol, E. Uçar, Ü. Atila, and M. Uçar, "An ensemble approach for classification of tympanic membrane conditions using soft voting classifier," *Multimed. Tools Appl.*, vol. 83, no. 32, pp. 77809–77830, Feb. 2024, doi: 10.1007/s11042-024-18631-z.

[22] L. Hermawan *et al.*, "Early Mental Health Detection with Machine Learning : A Practical Approach to Model Development and Implementation," *Indones. J. Electr. Eng. Informatics*, vol. 13, no. 2, pp. 509–524, Jun. 2025, doi: 10.52549/ijeei.v13i2.6111.

[23] J. Biswas, M. N. Hasan, M. S. R. Gazi, and M. M. Rahman, "Enhancing mental well-being: An artificial intelligence model for predicting mental disorders," *Array*, vol. 26, no. July, p. 100417, Jul. 2025, doi: 10.1016/j.array.2025.100417.

[24] H. G. Jabbar, "Advanced Threat Detection Using Soft and Hard Voting Techniques in Ensemble Learning," *J. Robot. Control*, vol. 5, no. 4, pp. 1104–1116, Jun. 2024, doi: 10.18196/JRC.V5I4.22005.

[25] S. Kudlapura Shivaiah, K. Krishnappa, N. Kumar Boraiah, P. Deepa Shenoy, and V. Kuppanna Rajuk, "Prediction of mental illness using ensemble model and grid search hyperparameter optimization," *Bull. Electr. Eng. Informatics*, vol. 13, no. 5, pp. 3726–3736, Oct. 2024, doi: 10.11591/eei.v13i5.7309.

[26] S. S. Shinde and A. S. Ghotkar, "From Questionnaires to Actionable Insights: Machine Learning for Mental Stress Detection," *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 6, pp. 29240–29250, Dec. 2025, doi: 10.48084/etasr.13513.

[27] T. ShamsEldin *et al.*, "Artificial intelligence for predicting depression anxiety and stress using psychometric data," *Sci. Rep.*, vol. 15, no. 1, p. 37282, Oct. 2025, doi: 10.1038/s41598-025-21301-1.

[28] A. Rai, "Explainable AI: from black box to glass box," *J. Acad. Mark. Sci.*, vol. 48, no. 1, pp. 137–141, Jan. 2020, doi: 10.1007/S11747-019-00710-5/TABLES/1.

[29] M. Ali, S. Ali, Q. Abbas, Z. Abbas, and S. W. Lee, "Artificial intelligence for mental health: A narrative review of applications, challenges, and future directions in digital health," *Digit. Heal.*, vol. 11, pp. 1–25, May 2025, doi: 10.1177/20552076251395548.

[30] P. Cruz-Gonzalez *et al.*, "Artificial intelligence in mental health care: a systematic review of diagnosis, monitoring, and intervention applications," *Psychol. Med.*, vol. 55, p. e18, Feb. 2025, doi: 10.1017/S0033291724003295.

[31] R. Tornero-Costa, A. Martinez-Millana, N. Azzopardi-Muscat, L. Lazeri, V. Traver, and D. Novillo-Ortiz, "Methodological and Quality Flaws in the Use of Artificial Intelligence in Mental Health Research: Systematic Review," *JMIR Ment. Heal.*, vol. 10, no. 1, p. e42045, Feb. 2023, doi: 10.2196/42045.

[32] H. Semmelrock *et al.*, "Reproducibility in machine-learning-based research: Overview, barriers, and drivers," *AI Mag.*, vol. 46, no. 2, p. e70002, Jun. 2025, doi: 10.1002/aaai.70002.

[33] S. Hameed, M. Nauman, N. Akhtar, M. A. B. Fayyaz, and R. Nawaz, "Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models," *Front. Artif. Intell.*, vol. 8, p. 1627078, Sep. 2025, doi: 10.3389/frai.2025.1627078.

[34] E. Kerz, S. Zanwar, Y. Qiao, and D. Wiechmann, "Toward explainable AI (XAI) for mental health detection based on language behavior," *Front. Psychiatry*, vol. 14, p. 1219479, Dec. 2023, doi: 10.3389/fpsyt.2023.1219479.

[35] E.-S. Atlam *et al.*, "Explainable artificial intelligence systems for predicting mental health problems in autistics," *Alexandria Eng. J.*, vol. 117, no. April, pp. 376–390, Apr. 2025, doi: 10.1016/j.aej.2024.12.120.

[36] J. J. Wadden, "Defining the undefinable: the black box problem in healthcare artificial intelligence," *J. Med. Ethics*, vol. 48, no. 10, pp. 764–768, Oct. 2022, doi: 10.1136/medethics-2021-107529.

[37] N. Thalpage, "Unlocking the Black Box: Explainable Artificial Intelligence (XAI) for Trust and Transparency in AI Systems," *J. Digit. Art Humanit.*, vol. 4, no. 1, pp. 31–36, Jun. 2023, doi: 10.33847/2712-8148.4.1_4.

[38] S. Hur *et al.*, "Comparison of SHAP and clinician friendly explanations reveals effects on clinical decision behaviour," *npj Digit. Med.*, vol. 8, no. 1, p. 578, Sep. 2025, doi: 10.1038/s41746-025-01958-8.

[39] Z. Ding *et al.*, "Trade-offs between machine learning and deep learning for mental illness detection on social media," *Sci. Rep.*, vol. 15, no. 1, p. 14497, Apr. 2025, doi: 10.1038/s41598-025-99167-6.

[40] J. Chung and J. Teo, "Single classifier vs. ensemble machine learning approaches for mental health prediction," *Brain Informatics*, vol. 10, no. 1, p. 1, Dec. 2023, doi: 10.1186/s40708-022-00180-6.

[41] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Comput. Sci.*, vol. 181, pp. 526–534, Jan. 2021, doi: 10.1016/j.procs.2021.01.199.

[42] P. F. Lovibond and S. H. Lovibond, "The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories," *Behav. Res. Ther.*, vol. 33, no. 3, pp. 335–343, Mar. 1995, doi: 10.1016/0005-7967(94)00075-U.

[43] V. Berisha *et al.*, "Digital medicine and the curse of dimensionality," *npj Digit. Med.*, vol. 4, no. 1, p. 153, Oct. 2021, doi: 10.1038/s41746-021-00521-5.

[44] A. Rácz, D. Bajusz, and K. Héberger, "Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification," *Molecules*, vol. 26, no. 4, p. 1111, Feb. 2021, doi: 10.3390/molecules26041111.

[45] M. T R, V. K. V, D. K. V, O. Geman, M. Margala, and M. Guduri, "The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification," *Healthc. Anal.*, vol. 4, no. December, p. 100247, Dec. 2023, doi: 10.1016/j.health.2023.100247.

[46] G. Van den Broeck, A. Lykov, M. Schleich, and D. Suciu, "On the Tractability of SHAP Explanations," *J. Artif. Intell. Res.*, vol. 74, pp. 851–886, Jun. 2022, doi: 10.1613/jair.1.13283.

[47] A. M. Salih *et al.*, "A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME," *Adv. Intell. Syst.*, vol. 7, no. 1, p. 2400304, Jan. 2025, doi: 10.1002/aisy.202400304.