Ingénierie des Systèmes d'Information
Vol. 31, No. 10, October, 2026, pp. 3283-3296
Journal homepage: http://iieta.org/journals/isi

An Explainable Two-Stage Soft-Voting Ensemble for Classifying Student Mental Health Severity and Follow-Up Recommendations

Pandu Pratama Putra[1*], Sarjon Defit[2], Gunadi Widi Nurcahyo[3]

[1] Department of Informatics Engineering, Universitas Lancang Kuning, Pekanbaru 28265, Indonesia
[2] Department of Information Technology, Universitas Putra Indonesia YPTK Padang, Padang 25145, Indonesia
[3] Department of Computer Science, Universitas Putra Indonesia YPTK Padang, Padang 25145, Indonesia

Corresponding Author Email: pandupratamaputra@unilak.ac.id

Copyright: ©2026 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/isi.311004

Received: 12 June 2026
Revised: 29 August 2026
Accepted: 14 September 2026
Available online: 31 October 2026

Keywords:
student mental health, mental health screening, soft-voting ensemble, explainable artificial intelligence, Local Interpretable Model-Agnostic Explanations, follow-up recommendation classification

ABSTRACT

Early screening of student mental health concerns requires models that can classify symptom severity while providing interpretable evidence for subsequent follow-up. This study presents an explainable, two-stage soft-voting ensemble for classifying depression, anxiety, and stress severity and assigning follow-up recommendation categories. The dataset comprises 960 free-text questionnaire responses developed around the three dimensions of the Depression Anxiety Stress Scales (DASS). Psychology experts reviewed and validated the severity labels and two follow-up categories: Psychological Counseling and Clinical Referral. Responses were preprocessed and represented using term frequency–inverse document frequency (TF-IDF). In Stage 1, Random Forest, C4.5 Decision Tree, and Multinomial Naïve Bayes probability estimates were combined with equal weights to predict five severity levels for each dimension. Under stratified five-fold cross-validation, the ensemble achieved accuracies of 89.79% for depression, 87.64% for anxiety, and 88.29% for stress, exceeding the corresponding individual classifiers. Out-of-fold severity predictions were then used as inputs to Stage 2, which classified follow-up recommendations with 91.88% accuracy, 92% precision, 90% recall, and a 91% F1-score. Local Interpretable Model-Agnostic Explanations (LIME) were applied at both stages to identify influential textual features and explain the contribution of the predicted severity profile to follow-up classification. The findings demonstrate the feasibility of integrating severity classification and follow-up categorization within a single interpretable pipeline. However, the evaluation was based on one dataset with expert-validated labels. External validation is needed to establish generalizability, and the reported classification metrics do not establish clinical validity. The framework should therefore be regarded as a preliminary screening and decision-support approach, not a diagnostic tool or a substitute for professional assessment.

1. INTRODUCTION

Mental health has become an increasingly important concern in educational environments, particularly among students who are exposed to academic pressure, performance expectations, and complex social interactions [1-4]. Depression, anxiety, and stress may affect academic performance, learning motivation, interpersonal relationships, and overall well-being. Consequently, early identification of psychological distress is important for supporting timely follow-up and promoting a more responsive educational environment [5]. A recent systematic review of machine learning approaches for detecting depression, anxiety, and stress among undergraduate students identified substantial research interest in computational mental health assessment, although concerns remain regarding internal validation, methodological heterogeneity, and the generalizability of reported performance [6].

Machine learning has increasingly been investigated as a computational approach for supporting mental health assessment. In student populations, previous research has demonstrated the potential of machine learning to identify psychological conditions from behavioral, demographic, physiological, and digital data [1, 6]. Iparraguirre-Villanueva et al. [1], for example, compared logistic regression, K-nearest neighbor, and decision tree models for classifying and predicting depression using a dataset of 787 college students. These findings demonstrate that machine learning can identify patterns associated with psychological conditions; however, predictive performance alone does not ensure that the resulting model is sufficiently transparent for mental-health-related decision support.

This limitation has motivated increasing interest in explainable artificial intelligence (XAI) for mental health applications. Recent studies have incorporated explainability techniques to clarify the factors underlying predictions of depression and other psychological conditions. Ahmed et al. [7] demonstrated that explainability methods such as SHapley

Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) can provide additional insight into the factors contributing to depression detection and severity classification. In text-based mental health analysis, Islam et al. [8] integrated ensemble learning with post-hoc explanations for depression-related emotion and severity detection, demonstrating the relevance of combining predictive performance with interpretable linguistic evidence. These studies indicate that explainability is particularly important when computational predictions concern psychologically meaningful outcomes.

Another challenge is that mental health classification frequently terminates at the prediction of a condition or severity level. However, initial screening in practice may also require information that can support subsequent professional follow-up. Retnaswamy et al. [2] explored machine-learning-based mental health treatment prediction, while Kelly et al. [9] developed an interpretable hierarchical model for mental health treatment prediction and emphasized the role of transparency when computational outputs are used to support professional assessment. These studies suggest that the relationship between symptom assessment and subsequent follow-up deserves explicit modeling rather than treating severity classification as an isolated final output.

From a methodological perspective, heterogeneous ensemble learning offers a mechanism for integrating classifiers with different learning characteristics. Random Forest can model complex nonlinear relationships, C4.5 Decision Tree provides a relatively interpretable rule-based structure, and Multinomial Naïve Bayes is computationally efficient for high-dimensional textual representations [10-12]. Previous studies in general text-classification domains have also demonstrated the ability of stacking, boosting, and voting strategies to combine complementary classifier outputs [13-16]. These general-domain studies are used in this study specifically to support the technical rationale for ensemble construction rather than as evidence of clinical or psychological applicability.

Despite these developments, several gaps remain within the scope addressed by this study. First, many machine learning approaches for student mental health focus primarily on predicting a mental health state or severity category [1, 6]. Second, studies incorporating explainability commonly explain a single prediction stage, whereas the relationship between textual indicators, multidimensional severity assessment, and subsequent follow-up classification remains less explicitly represented [7, 8]. Third, when depression, anxiety, and stress are evaluated simultaneously, their combined predicted profile can provide an informative intermediate representation before a follow-up category is generated. Finally, mental health decision-support applications require clear separation between computational recommendations and professional clinical judgment [9].

A two-stage architecture is therefore considered in this study to separate two related but conceptually different classification tasks. The first stage addresses the severity of depression, anxiety, and stress indicated by respondents' textual responses. The second stage considers the resulting multidimensional severity profile and classifies it into a predefined follow-up category. Separating these stages maintains an observable intermediate severity profile and allows the predictive behavior of each stage to be evaluated independently. However, because the second stage depends on the predictions generated by the first stage, errors in severity classification may propagate into the subsequent follow-up classification. This dependency is therefore explicitly considered in the experimental evaluation.

Explainability is incorporated at both stages to improve the transparency of the prediction pathway. LIME provides local explanations through an interpretable surrogate model fitted around an individual prediction. Applications of LIME in mental-health-related prediction, including severity assessment and text-based depression analysis, have been reported in previous studies [7, 8, 17]. In a related student-support context, Badrani et al. [18] explored SHAP and LIME for a hybrid neural network–SVM chatbot for student orientation in Morocco. This educational application should not be taken as evidence of validity for mental health screening or Clinical Referral. In the first stage of the proposed framework, LIME identifies textual features that support or oppose the predicted depression, anxiety, and stress severity categories. In the second stage, LIME explains how the predicted severity dimensions contribute to the resulting follow-up classification.

Based on these considerations, this study proposes an explainable two-stage heterogeneous soft voting ensemble framework for student mental health severity assessment and follow-up recommendation classification. Essay-based questionnaire responses developed around the depression, anxiety, and stress dimensions of the Depression Anxiety Stress Scales (DASS) are transformed using term frequency–inverse document frequency (TF-IDF) and classified using Random Forest, C4.5 Decision Tree, and Multinomial Naïve Bayes. Their probability outputs are integrated using equal-weight Soft Voting. The predicted depression, anxiety, and stress levels are subsequently combined into an intermediate severity profile and processed in a second classification stage to identify one of two expert-validated follow-up categories: Psychological Counseling or Clinical Referral.

The methodological contribution of this study does not lie in proposing a new individual classification algorithm. Instead, it lies in the structured integration of heterogeneous ensemble learning, multidimensional severity assessment, expert-validated follow-up classification, and local explainability within a unified two-stage architecture. This structure explicitly preserves the intermediate severity profile and provides explanations at both the text-to-severity and severity-to-follow-up levels. The framework is intended as a computational tool for initial screening and decision support and does not provide clinical diagnoses or replace assessment by qualified mental health professionals.

2. METHOD

This study employs a two-stage explainable machine learning framework to classify student mental health severity and categorize appropriate follow-up recommendations. The experimental process begins with 960 essay-based questionnaire responses developed around the depression, anxiety, and stress dimensions of the DASS. The severity labels and follow-up recommendation categories associated with the responses were established through consultation with and validation by psychology experts. The textual responses were then prepared through text cleaning, lowercasing, tokenization, stopword removal, stemming, and normalization, followed by TF-IDF transformation to obtain numerical feature representations.

The first classification stage independently estimates the severity of depression, anxiety, and stress from the processed textual responses. Three heterogeneous classifiers, namely Random Forest, C4.5 Decision Tree, and Multinomial Naïve Bayes, generate class-probability estimates that are aggregated using an equal-weight soft voting ensemble. Each mental health dimension is classified into one of five severity levels: Normal, Mild, Moderate, Severe, or Extremely Severe. LIME are subsequently applied to selected predictions to identify the textual features that support or oppose the predicted severity class.

The three predicted severity levels are then combined into an intermediate severity profile consisting of depression, anxiety, and stress levels. This profile serves as the input to the second classification stage, which categorizes each case into one of two expert-validated follow-up categories: Psychological Counseling or Clinical Referral. The second stage employs the same heterogeneous classifier configuration and soft voting mechanism, while LIME is used to explain the contribution of each predicted severity dimension to the resulting follow-up classification.

The two-stage design enables the severity assessment and follow-up classification processes to be examined separately while maintaining an explicit connection between them. Because the second-stage prediction depends on the severity profile produced in the first stage, the performance of each stage is evaluated independently to identify the predictive characteristics and potential influence of first-stage classification errors on the subsequent follow-up classification. The proposed framework is intended to support initial screening and follow-up decision support in educational settings and does not constitute clinical diagnosis or replace assessment by qualified mental health professionals. Figure 1 illustrates the overall methodological workflow of the proposed framework.

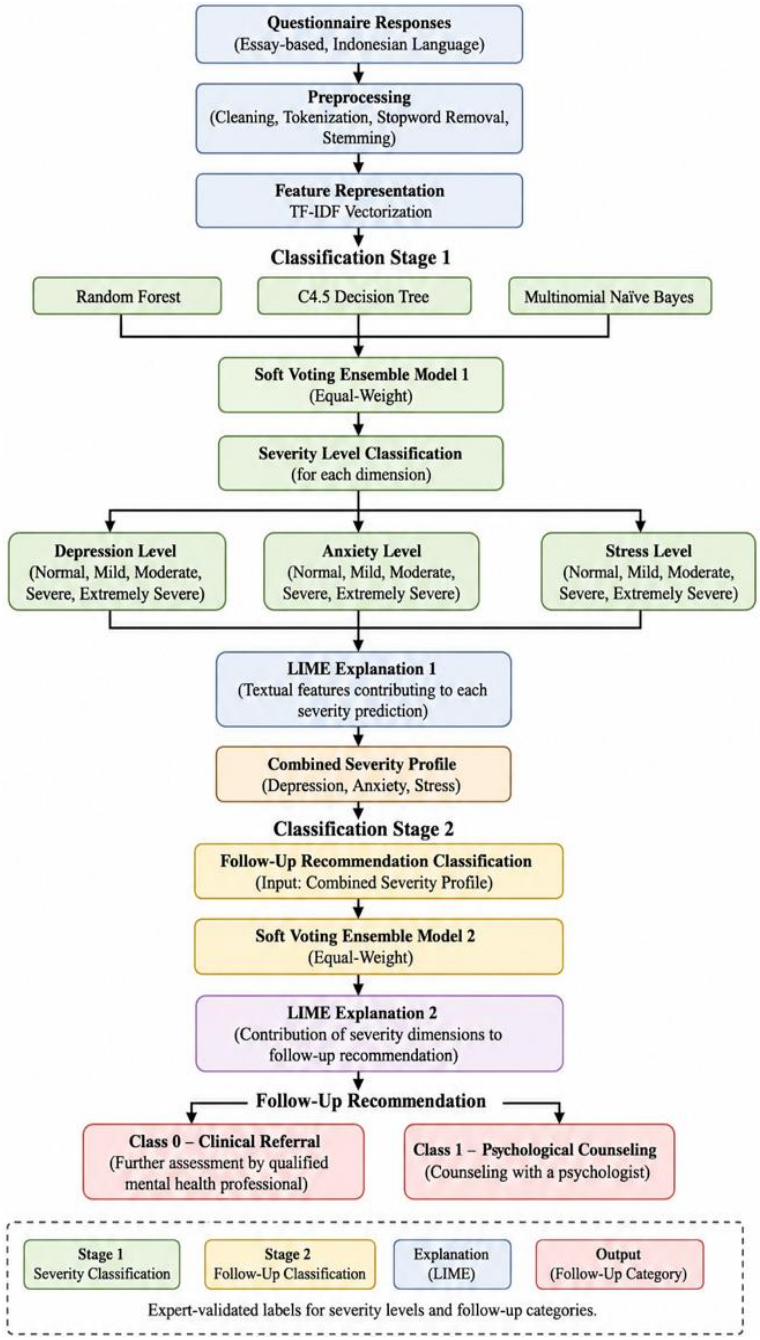

Figure 1. Methodology flow

2.1 Data collection

The dataset used in this study consists of 960 instances collected from respondents through an essay-based questionnaire developed around the depression, anxiety, and stress dimensions of the DASS. DASS is an established instrument for assessing depression, anxiety, and stress dimensions, and the Indonesian version of DASS-21 has also demonstrated satisfactory validity and reliability for assessing these psychological constructs [19]. However, unlike the original structured DASS format, the questionnaire used in this study elicits free-text responses developed around the corresponding psychological dimensions.

Because the questionnaire representation differs from the original structured instrument, severity labels were not generated automatically from textual features or directly inferred by the machine learning algorithms. The labeling procedure was conducted through consultation with and validation by psychology experts. The experts examined the respondents' responses within the conceptual dimensions of depression, anxiety, and stress and validated the corresponding severity categories. Each dimension was categorized into five severity levels: Normal, Mild, Moderate, Severe, and Extremely Severe.

In addition to the severity labels, follow-up recommendation labels were established through consultation with and validation by psychology experts. These labels consist of Psychological Counseling and Clinical Referral. Psychological Counseling represents cases for which counseling with a psychologist is considered an appropriate follow-up category, whereas Clinical Referral represents cases requiring further assessment by qualified mental health professionals. These categories serve as expert-validated decision-support labels and do not constitute clinical diagnoses, treatment prescriptions, or autonomous intervention decisions.

The resulting dataset therefore supports two interconnected classification tasks. The first task predicts depression, anxiety, and stress severity from textual responses, while the second task uses the combined predicted severity profile to classify an expert-validated follow-up category.

2.2 Data preprocessing

The responses collected from participants are expressed as free text and therefore cannot be processed directly by the selected classification algorithms. The raw responses undergo several preprocessing procedures before feature extraction. This stage aims to reduce textual noise, standardize linguistic variations, and preserve terms containing meaningful information associated with depression, anxiety, and stress [20].

The preprocessing process begins with text cleaning, during which punctuation marks, special symbols, redundant spaces, and irrelevant numerical characters are removed [21]. All textual data are subsequently converted into lowercase form to ensure that identical terms written using different capitalization styles are treated consistently. For example, the terms "Stress," "STRESS," and "stress" are represented as the same lexical item.

The cleaned responses are subsequently segmented into individual lexical units through word-level tokenization [22]. Frequently occurring words that provide limited discriminative information are removed through stopword removal [23]. Indonesian stopwords are removed using a predefined Indonesian stopword list, while terms considered relevant to psychological interpretation are retained.

Stemming is subsequently performed using the Sastrawi Indonesian stemming algorithm to transform morphological variations into their root forms [24]. This procedure reduces vocabulary variation and allows words sharing the same lexical root to be represented more consistently.

Text normalization is additionally applied because essay responses may contain informal expressions, abbreviations, spelling variations, and minor typographical errors. The normalization process converts such nonstandard expressions into standardized forms whenever appropriate. After preprocessing, the resulting corpus is transformed into numerical feature vectors using TF-IDF.

2.3 Term Frequency–Inverse Document Frequency

TF-IDF is employed to transform the preprocessed textual responses into weighted numerical representations [25-27]. TF-IDF assigns a weight to each term according to its frequency within a particular document and its distribution across the entire corpus.

Terms occurring frequently within a particular response but less frequently across the overall corpus receive relatively higher weights, whereas terms commonly appearing across many responses receive lower discriminative weights. The resulting representation enables the classification algorithms to recognize textual patterns associated with different levels of depression, anxiety, and stress.

Both unigram and bigram features are included by setting the n-gram range to (1, 2). Terms occurring in fewer than two documents are excluded using a minimum document frequency (min_df) of 2, while terms appearing in more than 95% of the documents are excluded using a maximum document frequency (max_df) of 0.95. The vocabulary is limited to a maximum of 5,000 features to control dimensionality relative to the available dataset size. Sublinear term-frequency scaling is applied, and the resulting vectors are normalized using L2 normalization.

The TF-IDF weight of term (i) in document (j) is calculated as follows:

$$w_{ij} = tf_{ij} x \log(N/df_i) \quad (1)$$

where,

w_{ij}: TF-IDF weight of term i in document j

tf_{ij}: Frequency of term i in document j

N: Total number of documents in the dataset

df_i: Number of documents containing term i

2.4 Base classifiers

This study employs three heterogeneous base classifiers: Random Forest, C4.5 Decision Tree, and Multinomial Naïve Bayes. These classifiers were selected because they represent distinct learning mechanisms that may provide complementary predictions within the ensemble framework.

Random Forest constructs multiple decision trees and aggregates their predictions, enabling the algorithm to capture complex relationships while reducing the tendency of an individual decision tree to overfit the training data [28]. C4.5 Decision Tree constructs hierarchical classification rules and provides a relatively interpretable tree-based decision

structure [29]. Multinomial Naïve Bayes employs a probabilistic classification mechanism and is particularly suitable for high-dimensional text representations such as TF-IDF [30].

The purpose of combining these classifiers is not to propose a new individual classification algorithm but to exploit their heterogeneous learning characteristics. Their probability outputs are therefore combined through a soft voting ensemble mechanism.

The selection of heterogeneous classifiers is also supported by previous research on computational mental health assessment. A systematic review of machine learning approaches for depression, anxiety, and stress among undergraduate students reported the use of various conventional machine learning algorithms and showed that predictive performance may differ depending on the target psychological condition, feature representation, and dataset characteristics [6]. This variation supports the use of classifiers with different learning mechanisms rather than assuming that a single algorithm will perform optimally across depression, anxiety, and stress classification tasks.

2.5 Heterogeneous soft voting ensemble model

The heterogeneous soft voting ensemble integrates Random Forest, C4.5 Decision Tree, and Multinomial Naïve Bayes within a unified probabilistic classification mechanism. Random Forest can capture complex relationships between features [31], C4.5 Decision Tree provides hierarchical decision structures [32], and Multinomial Naïve Bayes provides an efficient probabilistic approach for processing textual feature representations [33].

Unlike hard voting, which considers only the predicted class labels, Soft Voting incorporates the probability estimates generated by individual classifiers. Each classifier generates a probability distribution over the available classes, and these probabilities are subsequently aggregated [34-36].

In this study, all three classifiers are assigned equal weights, with each classifier contributing one-third of the aggregated probability. No weight optimization is performed. Equal weighting was selected to provide a direct evaluation of the complementary contribution of the heterogeneous classifiers without introducing an additional optimization procedure.

The application of ensemble learning in mental health text analysis is further supported by recent research demonstrating that probability-based ensemble strategies can combine complementary predictions for depression-related emotion and severity classification [7]. In the present study, the technical mechanism of probability aggregation follows established Soft Voting principles, while its effectiveness is specifically evaluated within the proposed student mental health severity and follow-up classification framework.

Let (x_i) represent the feature vector of the i-th instance, while (K) denotes the set of possible classes. The aggregated probability is calculated as follows:

$$P_{\text{ensemble}}(y = k|x) = \frac{\left[\begin{matrix} P_{\text{RF}}(y = k|x) + \\ P_{\text{C4.5}}(y = k|x) + P_{\text{MNB}}(y = k|x) \end{matrix}\right]}{3} \quad (2)$$

where,

$P_{\text{RF}}(y = k|x)$ = probability generated by Random Forest

$P_{\text{C4.5}}(y = k|x)$ = probability generated by C4.5 Decision Tree

$P_{\text{MNB}}(y = k|x)$ = probability generated by Multinomial Naïve Bayes

The final predicted class is determined as:

$$\hat{y} = \arg\max\ P_{\text{ensemble}}(y = k|x) \quad (3)$$

This mechanism incorporates the confidence information generated by all three classifiers rather than relying only on discrete class predictions. The same equal-weight soft voting principle is employed in both classification stages.

2.6 Classification stage 1: Mental health severity assessment

The first classification stage determines the severity levels of depression, anxiety, and stress from respondents' textual responses. Following preprocessing and TF-IDF transformation, the feature vectors are independently processed by Random Forest, C4.5 Decision Tree, and Multinomial Naïve Bayes.

Each classifier independently generates probability estimates for the available severity classes. These probabilities are subsequently aggregated using the equal-weight heterogeneous soft voting ensemble described in Section 2.5.

Three classification outputs are generated for each respondent: depression severity, anxiety severity, and stress severity. Each dimension is independently classified into one of five expert-validated severity categories: Normal, Mild, Moderate, Severe, or Extremely Severe.

The three predictions are subsequently combined into an intermediate severity profile:

$$Z_i = [D_i, A_i, S_i] \quad (4)$$

where,

D_i: predicted depression severity level

A_i: predicted anxiety severity level

S_i: predicted stress severity level

Each dimension is classified into one of five categories: Normal, Mild, Moderate, Severe, or Extremely Severe.

2.7 Local Interpretable Model-Agnostic Explanation-based explanation of severity predictions

To improve the transparency of the first classification stage, LIME is employed to explain individual severity predictions [37, 38]. LIME generates a locally interpretable approximation of the ensemble model around a particular instance and estimates the contribution of individual textual features to the resulting prediction.

The relevance of local explanation is particularly important in mental health applications, where predictive accuracy alone does not reveal why a model associates a particular response with a psychological severity category. Recent mental health studies have applied LIME and related XAI techniques to interpret depression detection and severity predictions [7], including linguistic feature contributions in text-based depression classification [8]. Accordingly, LIME is used in this study to identify the terms that support or oppose the predicted depression, anxiety, and stress severity levels.

The explanation is interpreted as evidence of model behavior rather than psychological causality. Representative correctly classified cases are examined for each mental health dimension, and the identified textual contributions are evaluated in relation to the psychological meaning of the

corresponding severity category.

2.8 Classification stage 2: Follow-up recommendation classification

The second classification stage uses the intermediate severity profile generated by the first stage as input for follow-up recommendation classification. The use of an intermediate psychological profile before generating a subsequent decision-support output is consistent with the broader objective of interpretable mental health decision-support models, in which symptom-related information is represented explicitly before subsequent classification or treatment-related predictions are produced [9].

The encoded depression, anxiety, and stress severity profile is processed using Random Forest, C4.5 Decision Tree, and Multinomial Naïve Bayes. Their probability estimates are subsequently integrated using the same equal-weight soft voting mechanism employed in the first stage. The resulting output consists of two expert-validated categories: Psychological Counseling and Clinical Referral.

Importantly, these categories represent follow-up classifications rather than autonomous clinical intervention decisions. Similar concerns regarding transparency and professional oversight have been emphasized in interpretable mental health decision-support research [9]. Therefore, the second-stage output is intended to support initial screening and professional follow-up rather than replace psychological or psychiatric assessment.

2.9 Local Interpretable Model-Agnostic Explanation-based explanation of follow-up recommendations

A second LIME analysis is applied to explain the classifications generated by the second-stage soft voting ensemble. In contrast to the first-stage explanation, which examines textual features, the second-stage explanation analyzes the encoded depression, anxiety, and stress severity profile.

The explanation identifies the contribution of each severity dimension to the predicted follow-up category. For example, an Extremely Severe depression prediction may provide a strong positive contribution toward Clinical Referral, whereas combinations of lower or moderate severity levels may contribute differently toward Psychological Counseling.

Representative cases from both follow-up categories are analyzed to demonstrate how different combinations of severity levels influence the resulting classification. Similar to the first-stage analysis, the selected cases are used to illustrate representative model behavior rather than being chosen solely on the basis of their explanation strength.

Applying LIME at both stages therefore provides two complementary forms of explanation. The first identifies how textual features influence depression, anxiety, and stress severity predictions, whereas the second identifies how the resulting severity profile contributes to the predicted follow-up category. This two-level explanation structure provides a transparent prediction pathway from textual responses to severity assessment and subsequently to follow-up classification.

2.10 Experimental setup and validation strategy

The experimental evaluation was conducted using stratified five-fold cross-validation on the complete dataset of 960 instances. The use of stratification was intended to preserve the distribution of the target classes as consistently as possible across the folds. A fixed random seed of 42 was applied to ensure reproducibility of the data partitioning.

In each cross-validation iteration, 80% of the dataset, corresponding to 768 instances, was used for model training, while the remaining 20%, corresponding to 192 instances, was used for testing. This procedure was repeated five times so that each instance was included in a test fold exactly once. Consequently, the final evaluation metrics and confusion matrices were calculated from the aggregated out-of-fold predictions of all 960 instances rather than from predictions generated on the training data.

All text preprocessing and TF-IDF transformations were fitted exclusively using the training portion of each fold and subsequently applied to the corresponding test fold. This procedure was employed to prevent information from the test fold from influencing vocabulary construction, TF-IDF weighting, or model training.

For the first classification stage, separate five-fold cross-validation procedures were conducted for depression, anxiety, and stress classification. In each fold, Random Forest, C4.5 Decision Tree, and Multinomial Naïve Bayes were trained using the training partition, and their probability estimates for the unseen test partition were combined using the equal-weight soft voting mechanism. Predictions obtained from the five test folds were subsequently concatenated to obtain out-of-fold severity predictions for all 960 instances.

The three out-of-fold predictions generated for depression, anxiety, and stress were then combined to construct an out-of-fold severity profile for each instance. Using out-of-fold predictions at this stage is important because each severity value used to construct the profile was generated by a first-stage model that had not been trained on the corresponding instance.

The resulting 960 out-of-fold severity profiles were used as the dataset for evaluating the second-stage follow-up recommendation classifier. A second stratified five-fold cross-validation procedure was applied to these profiles. In each fold, 768 severity profiles were used for training and 192 profiles were used for testing. Random Forest, C4.5 Decision Tree, and Multinomial Naïve Bayes were again combined using equal-weight soft voting. The predictions from the five test folds were aggregated so that the final second-stage evaluation also represented all 960 instances.

This evaluation strategy provides two levels of protection against overly optimistic performance estimates. First, the severity profile used in the second stage is constructed from out-of-fold Stage 1 predictions rather than predictions generated from instances already observed during Stage 1 training. Second, Stage 2 itself is evaluated on unseen folds through stratified cross-validation.

Accuracy, precision, recall, and F1-score were used as the primary evaluation metrics. Confusion matrices were additionally examined to identify class-specific prediction patterns. The same fold structure and evaluation procedure were applied when comparing the proposed heterogeneous soft voting ensemble with the individual base classifiers to ensure a consistent and fair comparison.

Given the relatively limited dataset size, five-fold stratified cross-validation was selected to allow all 960 instances to contribute to both model development and out-of-sample evaluation while ensuring that every reported prediction was

generated when the corresponding instance was excluded from the relevant model-training fold.

2.11 Consistency assessment of model explanations

To complement the illustrative LIME visualizations, the interpretability analysis considers the psychological relevance of the features identified by the explanation method. The interpretation of selected LIME outputs is examined in relation to the conceptual characteristics of depression, anxiety, and stress used during expert validation of the dataset.

Particular attention is given to whether positively contributing textual features are psychologically consistent with the predicted severity dimension and whether the direction of their contribution remains conceptually reasonable across similar responses. For the second-stage analysis, the contribution of depression, anxiety, and stress severity levels is examined in relation to the expert-validated follow-up categories.

This additional assessment is intended to prevent the explainability analysis from relying exclusively on individual visualization examples. Nevertheless, LIME remains a local post-hoc explanation technique, and its outputs should be interpreted as explanations of model behavior rather than direct evidence of psychological causality.

3. RESULTS AND DISCUSSION

This section presents the experimental results of the proposed two-stage heterogeneous soft voting ensemble framework. All reported performance values were obtained from aggregated out-of-fold predictions generated through stratified five-fold cross-validation. Therefore, although the complete dataset contains 960 instances, each individual prediction was generated by a model that did not use the corresponding instance during training.

The evaluation is organized into two stages. Stage 1 assesses depression, anxiety, and stress severity classification from textual responses, while Stage 2 evaluates expert-validated follow-up recommendation classification using the out-of-fold severity profiles generated by Stage 1. Accuracy, precision, recall, and F1-score are reported, while confusion matrices are constructed by aggregating the predictions obtained from the five test folds.

3.1 Mental health severity classification

The first stage independently classifies depression, anxiety, and stress into five severity categories: Normal, Mild, Moderate, Severe, and Extremely Severe. Separate stratified five-fold cross-validation procedures were conducted for the three mental health dimensions. In each fold, 768 instances were used for training and 192 unseen instances were used for testing. Predictions from all five test folds were then aggregated, producing one out-of-fold prediction for each of the 960 instances.

Figure 2 presents the aggregated performance of the proposed heterogeneous soft voting ensemble for depression severity classification.

Based on the aggregated out-of-fold predictions, the proposed model achieved an overall depression classification accuracy of 89.79%. The Normal category obtained a precision of 0.87, recall of 0.97, and F1-score of 0.92. The Mild category achieved a precision of 0.91 and recall of 0.86, while the Moderate category obtained a precision of 0.91, recall of 0.83, and an F1-score of 0.87.

The Severe category achieved a precision of 0.86, recall of 1.00, and an F1-score of 0.92. Meanwhile, the Extremely Severe category achieved a precision of 0.90, recall of 0.83, and F1-score of 0.86. Overall, the macro-average precision was 0.89, while the macro-average recall and F1-score were both 0.90.

Figure 3 presents the confusion matrix constructed from the aggregated out-of-fold predictions.

Across the five test folds, the model correctly classified 195 Normal, 172 Mild, 159 Moderate, 184 Severe, and 152 Extremely Severe instances. In total, 862 of the 960 out-of-fold predictions were correct, corresponding to an overall accuracy of 89.79%.

The confusion matrix therefore represents predictions generated for all 960 instances while ensuring that each prediction was obtained when the corresponding instance belonged to an unseen test fold. This distinction is important because it prevents the reported accuracy from representing training-set performance.

	precision	recall	f1-score	support
Normal	0.87	0.97	0.92	201
Mild	0.91	0.86	0.91	200
Moderate	0.91	0.83	0.87	192
Severe	0.86	1.00	0.92	184
Very Severe	0.90	0.83	0.86	183
accuracy			0.90	960
macro avg	0.89	0.90	0.90	960
weighted avg	0.90	0.90	0.90	960

Accuracy = 0.8979166666666667

Figure 2. Performance evaluation metrics for depression severity classification

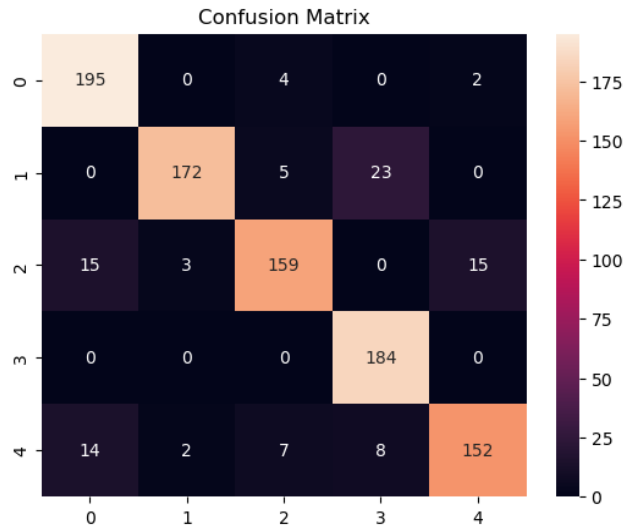

Figure 3. Confusion matrix for depression classification results

3.2 Comparison of severity classification models

Table 1 compares the proposed heterogeneous soft voting ensemble with the individual classifiers across depression, anxiety, and stress.

Table 1. Performance comparison for mental health severity classification

Model	Dimension	Accuracy	Precision	Recall	F1-Score
Proposed Model	Depression	89.79%	89%	90%	90%
	Anxiety	87.64%	88%	88%	88%
	Stress	88.29%	89%	88%	88%
Random Forest	Depression	85.65%	86%	86%	86%
	Anxiety	86.64%	87%	87%	87%
	Stress	85.29%	85%	85%	85%
C4.5	Depression	80.43%	80%	81%	80%
	Anxiety	79.94%	80%	80%	80%
	Stress	80.23%	81%	80%	80%
MNB	Depression	86.79%	87%	87%	87%
	Anxiety	87.43%	87%	87%	87%
	Stress	85.29%	85%	85%	85%

The proposed model achieved the highest accuracy across all three mental health dimensions. For depression, the ensemble achieved 89.79%, compared with 86.79% for Multinomial Naïve Bayes, 85.65% for Random Forest, and 80.43% for C4.5 Decision Tree. Relative to the strongest individual classifier, Multinomial Naïve Bayes, the ensemble increased depression classification accuracy by 3.00 percentage points.

For anxiety classification, the difference was substantially smaller. The proposed model achieved an accuracy of 87.64%, compared with 87.43% for Multinomial Naïve Bayes and 86.64% for Random Forest. The improvement over the strongest individual classifier was therefore only 0.21 percentage points.

This finding is important because it indicates that ensemble learning does not provide the same magnitude of improvement for every mental health dimension. For anxiety, Multinomial Naïve Bayes already produced strong predictions, leaving relatively little room for improvement through probability aggregation. Consequently, the contribution of the ensemble should be interpreted as improving overall consistency rather than producing uniformly large improvements for every classification task.

For stress classification, the proposed model achieved 88.29% accuracy, compared with 85.29% for both Random Forest and Multinomial Naïve Bayes and 80.23% for C4.5 Decision Tree. This corresponds to an improvement of 3.00 percentage points compared with the strongest individual classifiers.

Across the three dimensions, the average accuracy of the proposed ensemble was 88.57%, compared with 86.50% for Multinomial Naïve Bayes, 85.86% for Random Forest, and 80.20% for C4.5 Decision Tree. The results therefore indicate that combining heterogeneous probability estimates provides more consistent performance across depression, anxiety, and stress than relying on one classifier.

The difference in performance among the three dimensions also reinforces the rationale for using heterogeneous classifiers. No individual model exhibits the same relative strength across all classification tasks. This observation is consistent with previous mental health machine learning research showing that model performance varies according to the psychological condition, feature representation, and dataset characteristics [6].

The purpose of this comparison is specifically to evaluate the contribution of the soft voting mechanism relative to its constituent classifiers. Therefore, these results demonstrate an internal ensemble benefit but should not be interpreted as evidence that the proposed approach outperforms all existing mental health text-classification architectures.

3.3 Local Interpretable Model-Agnostic Explanations explanation of severity predictions

In addition to predictive performance, LIME was applied to selected predictions to examine how textual features influenced the severity classifications. Explainability is particularly relevant in mental health applications because a predicted severity label alone does not reveal the textual evidence considered important by the model. Previous studies have similarly demonstrated the value of XAI techniques for interpreting mental health predictions and depression severity classification [7, 8].

Figure 4 presents a representative LIME explanation for depression severity.

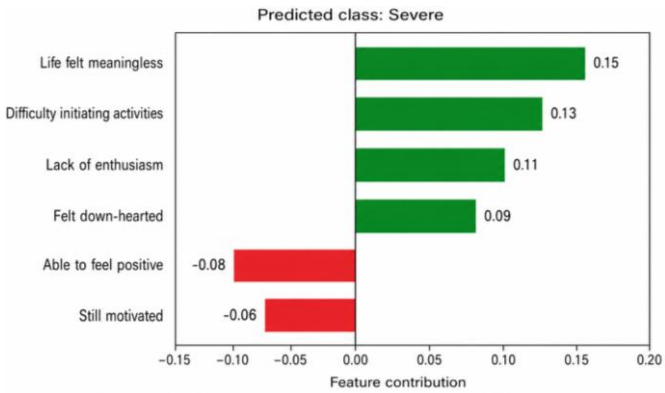

Figure 4. Local Interpretable Model-Agnostic Explanations (LIME)-based explanation of depression severity prediction

For the selected instance, the model predicted Severe depression. The strongest positive feature was "life felt meaningless," with a contribution value of 0.15, followed by "difficulty initiating activities" at 0.13, "lack of enthusiasm" at 0.11, and "felt down-hearted" at 0.09. Conversely, "able to feel positive" and "still motivated" contributed negatively, with values of −0.08 and −0.06, respectively.

The explanation indicates that the model associated expressions reflecting loss of meaning, reduced motivation, and negative emotional states with the predicted Severe category, while expressions reflecting positive affect or motivation reduced the probability of the prediction.

Figure 5 presents an explanation for an anxiety prediction.

For the selected Moderate anxiety prediction, the strongest positive contribution was generated by "felt scared without reason" at 0.14, followed by "panic feeling" at 0.12, "heart

racing" at 0.10, and "dry mouth" at 0.08. Conversely, "felt calm" and "no trembling" contributed negatively.

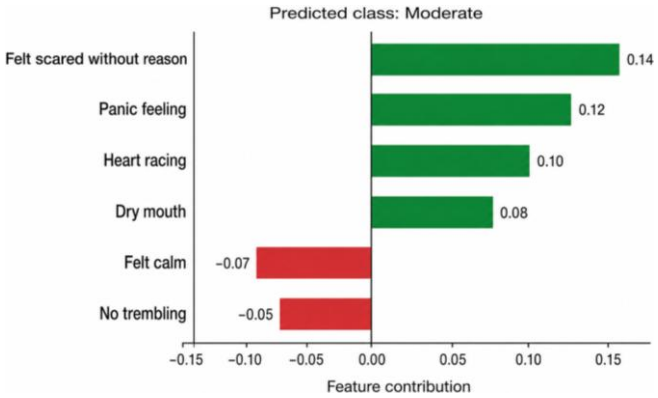

Figure 5. Local Interpretable Model-Agnostic Explanations (LIME)-based explanation of anxiety severity prediction

The explanation suggests that the model relied on a combination of emotional and physical anxiety-related expressions. This illustrates how local explainability can reveal why a particular response is assigned to a severity category rather than presenting only the final prediction.

Figure 6 presents a representative explanation for stress severity.

For the selected Moderate stress prediction, "difficulty relaxing" provided the strongest positive contribution at 0.14, followed by "over-reacting" at 0.11, "easily agitated" at 0.09, and "impatience" at 0.07. In contrast, "felt relaxed" and "rarely irritated" contributed negatively.

Across the three examples, the direction of the LIME contributions is conceptually consistent with the corresponding severity dimensions. However, these explanations should be interpreted as local explanations of model behavior rather than evidence that the identified words have a causal psychological relationship with depression, anxiety, or stress. This distinction is particularly important when XAI methods are applied in mental-health-related decision-support contexts [7, 8].

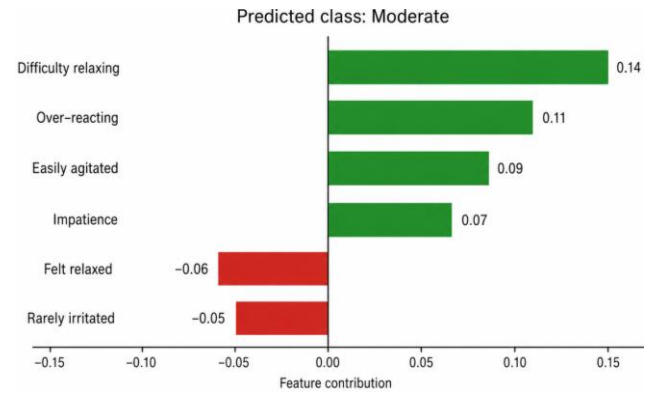

Figure 6. Local Interpretable Model-Agnostic Explanations (LIME)-based explanation of stress severity prediction

3.4 Follow-up recommendation classification

The second stage evaluates the ability of the proposed framework to classify follow-up recommendations using the combined depression, anxiety, and stress severity predictions generated in Stage 1. Each respondent is represented by a severity profile comprising three predicted values, corresponding to the three mental health dimensions.

The severity profiles were constructed from the out-of-fold predictions obtained during Stage 1 five-fold cross-validation. Each severity prediction was therefore generated by a model that had not been trained on the corresponding respondent's data. The three predictions were subsequently combined to produce one severity profile per respondent.

A total of 960 severity profiles were obtained. According to the expert-validated follow-up labels, 580 instances (60.42%) belonged to the Clinical Referral category, while 380 instances (39.58%) belonged to the Psychological Counseling category.

The second-stage classifier was evaluated using stratified five-fold cross-validation. In each iteration, 768 severity profiles were used for training and 192 profiles were used for testing. Each profile appeared in a test fold exactly once, and predictions from all five test folds were aggregated to calculate the reported evaluation metrics.

Figure 7 presents the classification report for the proposed heterogeneous soft voting ensemble. Based on the aggregated predictions, the model correctly classified 882 of the 960 instances, achieving an overall accuracy of 91.88%.

For the Clinical Referral category, the model achieved a precision of 0.90, a recall of 0.97, and an F1-score of 0.94. For the Psychological Counseling category, precision was 0.95, recall was 0.84, and the F1-score was 0.89. The macro-average precision, recall, and F1-score were 0.93, 0.91, and 0.91, respectively. The weighted-average precision, recall, and F1-score were all 0.92. All metrics were calculated from the underlying prediction counts before rounding to two decimal places.

The results show different classification patterns for the two recommendation categories. Clinical Referral achieved higher recall, indicating that the model identified most instances assigned to this category in the reference labels. Psychological Counseling achieved higher precision but lower recall, indicating that predictions assigned to this category were generally correct, although a substantial proportion of its actual instances were classified as Clinical Referral.

Figure 8 presents the confusion matrix derived from the aggregated predictions and provides further detail on these errors. Of the 580 Clinical Referral instances, 563 were correctly classified and 17 were incorrectly classified as Psychological Counseling. Of the 380 Psychological Counseling instances, 319 were correctly classified and 61 were incorrectly classified as Clinical Referral.

	precision	recall	f1-score	support
0	0.90	0.97	0.94	580
1	0.95	0.84	0.89	380
accuracy			0.92	960
macro avg	0.93	0.91	0.91	960
weighted avg	0.92	0.92	0.92	960

Accuracy : 0.91875

Figure 7. Performance evaluation metrics for follow-up recommendation classification

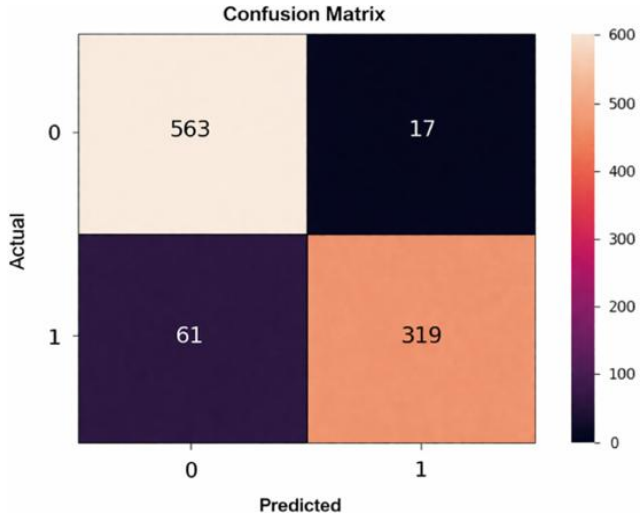

Figure 8. Confusion matrix for follow-up recommendation classification based on aggregated out-of-fold predictions

Overall, the model produced 78 classification errors. Of these, 61 occurred when Psychological Counseling instances were classified as Clinical Referral, while 17 occurred in the opposite direction. This asymmetry explains the higher recall for Clinical Referral and the lower recall for Psychological Counseling. These findings reflect the agreement between the model's predictions and the expert-validated recommendation labels within the evaluated dataset.

3.5 Comparison of follow-up classification models

Table 2 compares the performance of the proposed ensemble and its individual base classifiers for follow-up recommendation classification. The proposed ensemble achieved the highest reported accuracy, precision, recall, and F1-score among the evaluated models.

Among the individual classifiers, Random Forest produced the strongest performance, achieving an accuracy of 90.12%, a precision of 90%, a recall of 89%, and an F1-score of 90%. C4.5 Decision Tree achieved an accuracy of 89.43%, with precision, recall, and F1-score values of 89%. Meanwhile, Multinomial Naïve Bayes obtained the lowest performance, with an accuracy of 87.65% and precision, recall, and F1-score values of 87%.

Table 2. Performance comparison for follow-up recommendation classification

Model	Accuracy	Precision	Recall	F1-Score
Proposed Model	91.88%	92%	90%	91%
Random Forest	90.12%	90%	89%	90%
C4.5	89.43%	89%	89%	89%
Multinomial Naïve Bayes	87.65%	87%	87%	87%

Compared with Random Forest as the strongest individual classifier, the proposed model improved accuracy by 1.76 percentage points. The proposed model also increased precision by 2 percentage points, recall by 1 percentage point, and F1-score by 1 percentage point. These results indicate that combining Random Forest, C4.5 Decision Tree, and Multinomial Naïve Bayes through the soft voting mechanism provides more consistent performance than the use of individual classifiers for follow-up recommendation classification.

3.6 Local Interpretable Model-Agnostic Explanations explanation of follow-up classification

LIME was subsequently applied to the second-stage model to examine how the predicted depression, anxiety, and stress severity levels contributed to each follow-up classification.

Unlike the first-stage explanations, which operate on TF-IDF textual features, the second-stage explanations operate on the encoded severity profile. Consequently, the resulting explanation provides a direct interpretation of how the three predicted mental health dimensions influence the follow-up category.

Figure 9 presents a representative Psychological Counseling prediction.

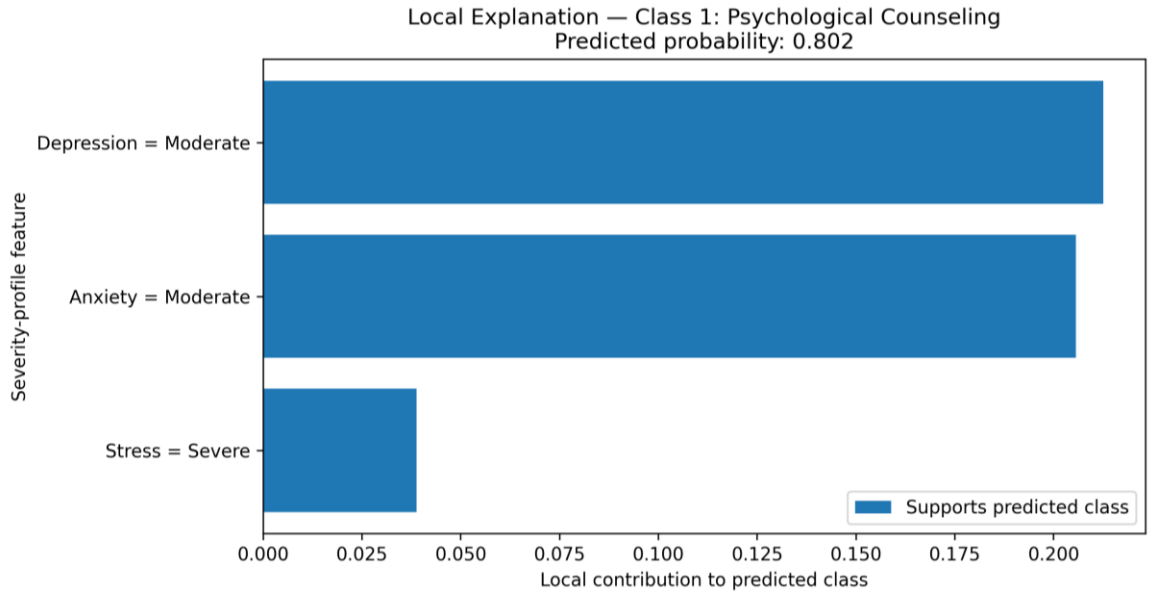

Figure 9. Local Interpretable Model-Agnostic Explanations (LIME)-based local explanation of Psychological Counseling classification

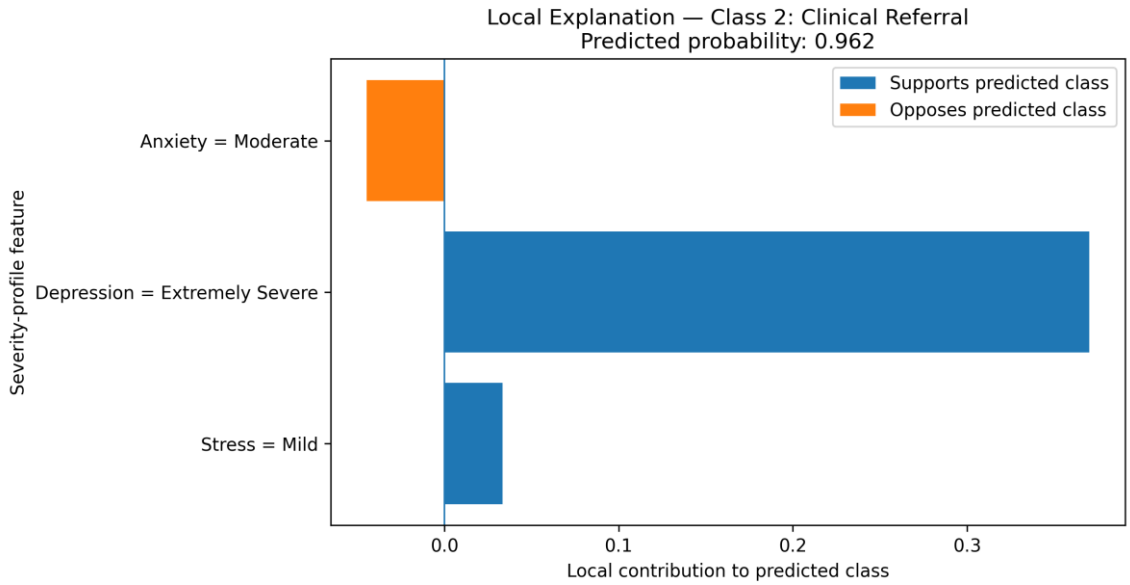

Figure 10. Local Interpretable Model-Agnostic Explanations (LIME)-based local explanation of Clinical Referral classification

For the selected respondent, the model predicted Psychological Counseling with a probability of 0.802. Depression = Moderate produced the largest positive local contribution, followed by Anxiety = Moderate. Stress = Severe also contributed positively but with a smaller effect.

The explanation indicates that the prediction was influenced primarily by the combination of moderate depression and anxiety levels rather than by a single severity dimension. This demonstrates the role of the intermediate severity profile in Stage 2, where multiple mental health dimensions can contribute simultaneously to the resulting classification.

Figure 10 presents a representative Clinical Referral prediction.

For this instance, the model predicted Clinical Referral with a probability of 0.962. Depression = Extremely Severe generated the largest positive contribution toward the predicted category. Stress = Mild also contributed positively but with a substantially smaller effect, whereas Anxiety = Moderate contributed negatively.

The explanation indicates that the Clinical Referral prediction was predominantly influenced by the Extremely Severe depression level for this particular instance. Importantly, the result should be interpreted as an explanation of how the trained model reproduced the expert-validated follow-up label rather than as evidence that a particular severity level automatically requires a specific clinical intervention.

The two Stage 2 examples demonstrate that LIME can expose the relative contribution of different severity dimensions instead of presenting the final follow-up category as an unexplained prediction. This form of interpretability is consistent with the broader objective of explainable mental-health decision-support systems, where transparency regarding the factors underlying predictions is important [9].

3.7 Overall discussion

Taken together, the experimental results demonstrate three main characteristics of the proposed framework.

First, heterogeneous soft voting produced more consistent classification performance than the individual Random Forest, C4.5 Decision Tree, and Multinomial Naïve Bayes models across the evaluated tasks. The largest improvements were observed for depression and stress, whereas the improvement for anxiety was comparatively small. This indicates that the advantage of ensemble learning depends on the complementary behavior of the constituent classifiers rather than guaranteeing a uniform increase for every prediction task.

Second, the two-stage architecture provides an explicit intermediate representation of depression, anxiety, and stress severity before follow-up classification. This differs conceptually from a single-stage architecture that maps textual input directly to a final follow-up output. Maintaining the severity profile allows the individual psychological dimensions to remain observable and enables separate evaluation of Stage 1 and Stage 2. This transparency is particularly relevant to interpretable mental health decision-support research [9].

Third, LIME provides complementary explanations at two levels. The first stage explains which textual features influence individual severity predictions, while the second stage explains how the resulting depression, anxiety, and stress levels contribute to the follow-up classification. Previous XAI research in mental health has similarly emphasized the value of explaining the factors underlying depression and severity predictions [7, 8]. The present framework extends this interpretability across two connected prediction stages.

Nevertheless, the results should be interpreted within the scope of the evaluated dataset. The proposed system learns patterns associated with expert-validated severity and follow-up labels and is intended to support initial screening. Its outputs do not establish psychological diagnoses or determine clinical interventions. Professional interpretation remains necessary, particularly because classification errors in Stage 1 may propagate to Stage 2 and because local XAI explanations describe model behavior rather than psychological causality.

4. CONCLUSIONS

This study developed an explainable two-stage heterogeneous soft voting ensemble framework for student mental health severity assessment and follow-up recommendation classification. The proposed framework

integrates Random Forest, C4.5 Decision Tree, and Multinomial Naïve Bayes using an equal-weight soft voting mechanism. In the first stage, essay-based questionnaire responses were classified independently into depression, anxiety, and stress severity levels. Based on aggregated out-of-fold predictions obtained through stratified five-fold cross-validation, the proposed model achieved accuracies of 89.79% for depression, 87.64% for anxiety, and 88.29% for stress. Across the three mental health dimensions, the ensemble consistently outperformed the individual base classifiers, indicating that combining classifiers with heterogeneous learning characteristics can provide more stable predictive performance within the evaluated dataset.

The predicted depression, anxiety, and stress severity levels were subsequently combined into an intermediate severity profile and used as input for the second-stage follow-up recommendation classification. The follow-up categories, Psychological Counseling and Clinical Referral, were established through consultation with and validation by psychology experts. Using the 960 out-of-fold severity profiles, the proposed model achieved an accuracy of 91.88%, with a precision of 92%, a recall of 90%, and an F1-score of 91%. These results indicate that the combined severity profile provides useful information for distinguishing between the two expert-validated follow-up categories. However, the second-stage output should be interpreted as a classification of expert-validated follow-up labels rather than as an autonomous clinical intervention decision.

To improve transparency, LIME was applied at both classification stages. In the first stage, LIME identified textual features that contributed to the predicted depression, anxiety, and stress severity levels. In the second stage, LIME clarified how the predicted severity dimensions contributed to the follow-up classification. The explanation results showed that severe or extremely severe depression levels may provide a strong contribution to Clinical Referral predictions in particular instances, while combinations of moderate depression and anxiety levels may contribute to Psychological Counseling predictions. These explanations should be interpreted as local representations of model behavior rather than as evidence of psychological causality.

The main contribution of this study lies in the integration of heterogeneous ensemble learning, multidimensional mental health severity assessment, expert-validated follow-up classification, and local explainability within a unified two-stage architecture. The intermediate severity profile enables the prediction process to remain observable between the text-based assessment stage and the subsequent follow-up classification stage.

Several limitations should be considered when interpreting the findings. The dataset consists of 960 instances, which remains relatively limited for five-class severity classification and may not fully represent the linguistic and psychological diversity of broader student populations. In addition, because the second-stage classifier depends on severity predictions generated in the first stage, errors in depression, anxiety, or stress classification may propagate to the subsequent follow-up classification. Although stratified five-fold cross-validation provides internal validation across the complete dataset, external validation is still required before the framework can be generalized to other educational populations.

Future research should therefore evaluate the proposed framework using larger and more diverse datasets collected from multiple educational institutions, conduct external validation, and involve additional qualified psychology or mental health experts in validating the target labels and model explanations. Comparisons with additional conventional machine learning, deep learning, and transformer-based text classification approaches should also be considered to establish the competitiveness of the proposed framework beyond its constituent classifiers. Further investigation may additionally explore model calibration, optimized ensemble weighting, and strategies for reducing error propagation between the two classification stages. The proposed framework is intended to support initial screening and follow-up decision support in educational settings and does not replace psychological assessment, clinical diagnosis, or professional judgment by qualified mental health professionals.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Universitas Putra Indonesia YPTK Padang and Universitas Lancang Kuning for their support and academic contributions to this research. This study is part of a doctoral dissertation conducted in the Doctoral Program in Information Technology at Universitas Putra Indonesia YPTK Padang. The authors also appreciate the valuable guidance, institutional support, and academic environment provided throughout the completion of this research.

REFERENCES

[1] Iparraguirre-Villanueva, O., Paulino-Moreno, C., Epifanía-Huerta, A., Torres-Ceclén, C. (2024). Machine learning models to classify and predict depression in college students. International Journal of Interactive Mobile Technologies, 18(14): 148-163. https://doi.org/10.3991/ijim.v18i14.48669

[2] Retnaswamy, C., Muthu, A.B.A., Duraipandi, B., Manickam, R. (2025). Mental health treatment prediction using machine learning. Biomedical and Pharmacology Journal, 18(3): 1925-1937. https://doi.org/10.13005/bpj/3225

[3] Srivani, U., Rakshitha, S., Kumar, B.P., Priyanka, M., Raj, N.S.Y., Anuradha, K. (2026). Machine learning models based mental health detection. International Journal of Marketing Management, 14(1): 118-127.

[4] Anderson, T.L., Valiauga, R., Tallo, C., et al. (2025). Contributing factors to the rise in adolescent anxiety and associated mental health disorders: A narrative review of current literature. Journal of Child and Adolescent Psychiatric Nursing, 38(1): e70009. https://doi.org/10.1111/jcap.70009

[5] Colizzi, M., Lasalvia, A., Ruggeri, M. (2020). Prevention and early intervention in youth mental health: Is it time for a multidisciplinary and trans-diagnostic model for care? International Journal of Mental Health Systems, 14(1): 23. https://doi.org/10.1186/s13033-020-00356-9

[6] Schaab, B.L., Calvetti, P.Ü., Hoffmann, S., et al. (2024). How do machine learning models perform in the detection of depression, anxiety, and stress among undergraduate students? A systematic review. Cadernos de Saúde Pública, 40: e00029323. https://doi.org/10.1590/0102-311XEN029323

[7] Ahmed, I., Brahmacharimayum, A., Ali, R.H., Khan, T.A., Ahmad, M.O. (2025). Explainable AI for depression detection and severity classification from activity data: Development and evaluation study of an interpretable framework. JMIR Mental Health, 12(1): e72038. https://doi.org/10.2196/72038
[8] Islam, S., Haque, R., Khan, M.A., et al. (2026). Ensemble transformer with post-hoc explanations for depression emotion and severity detection. iScience, 29(2): 114605. https://doi.org/10.1016/j.isci.2025.114605
[9] Kelly, A., Jensen, E.K., Grua, E.M., Mathiasen, K., Van de Ven, P. (2025). An interpretable model with probabilistic integrated scoring for mental health treatment prediction: Design study. JMIR Medical Informatics, 13: e64617. https://doi.org/10.2196/64617
[10] Sovia, R., Anam, M.K., Wisky, I.A., Permana, R., Rahmi, N.A., Zain, R.H. (2026). Robust predictive model for heart disease diagnosis using advanced machine learning techniques. Journal of Applied Data Sciences, 7(1): 541-553. https://doi.org/10.47738/jads.v7i1.1092
[11] Sudianto, S., Wahyuningtias, P., Utami, H.W., Raihan, U.A., Hanifah, H.N., Adanson, Y.N. (2022). Comparison of random forest and support vector machine methods on twitter sentiment analysis (Case study: Internet selebgram rachel vennya escape from quarantine). Jurnal Teknik Informatika, 3(1): 141-145. https://doi.org/10.20884/1.jutif.2022.3.1.168
[12] Singh, J., Tripathi, P. (2021). Sentiment analysis of twitter data by making use of SVM, random forest and decision tree algorithm. In 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, pp. 193-198. https://doi.org/10.1109/CSNT51715.2021.9509679
[13] Siregar, A.Y., Arifin, A.S. (2024). Enhancing XGBoost classification with SVM-SMOTE & EasyEnsemble for imbalanced telemedicine sentiment data. Jurnal Indonesia Sosial Teknologi, 5(10): 5465-5474. https://doi.org/10.59141/jist.v5i10.1160
[14] Santoso, D.B., Munna, A., Ningsih, D.H.U. (2024). Improved playstore review sentiment classification accuracy with stacking ensemble. Journal of Soft Computing Exploration, 5(1): 38-45. https://doi.org/10.52465/joscex.v5i1.247
[15] Nyandwi, V., Habimana, O., Enan, N.M. (2023). Ensemble machine learning-based sentiment analysis model for teachers' performance evaluation. International Journal of Advances in Engineering and Management, 5(4): 1220-1233. https://doi.org/10.35629/5252-050412201233
[16] Abbas, A.K., Salih, A.K., Hussein, H.A., Hussein, Q.M., Abdulwahhab, S.A. (2020). Twitter sentiment analysis using an ensemble majority vote classifier. Journal of Southwest Jiaotong University, 55(1). https://doi.org/10.35741/issn.0258-2724.55.1.9
[17] Laila, A.B.Z. (2026). Detecting the impact of social media on users' mental health using machine learning and XAI. Jurnal Buana Informatika, 17(1): 53-65. https://doi.org/10.24002/jbi.v17i1.13409
[18] Badrani, M., Marouan, A., Kannouf, N., Chetouani, A. (2026). An explainable chatbot for student orientation in Morocco: SHAP and LIME on a hybrid neural network-SVM model. International Journal of Engineering Pedagogy, 16(6): 164-175. https://doi.org/10.3991/ijep.v16i6.63255
[19] Ifdil, I., Syahputra, Y., Fadli, R.P., et al. (2020). The depression anxiety stress scales (DASS-21): An Indonesian validation measure of the depression anxiety stress. The International Journal of Counseling and Education, 5(5): 205-215. https://doi.org/10.23916/0020200536840
[20] Arsyah, U.I., Pratiwi, M., Fryonanda, H., Anam, M.K., Munawir, M. (2025). ACLM Model: A CNN-LSTM and machine learning approach for analyzing tourist satisfaction to improve priority tourism services. Journal of Applied Data Sciences, 6(4): 2891-2906. https://doi.org/10.47738/jads.v6i4.974
[21] Begum, S.G., Sree, P.K. (2023). Drug recommendations using a "reviews and sentiment analysis" by a recurrent neural network. Indonesian Journal of Multidisciplinary Science, 2(9): 3085-3094. https://doi.org/10.55324/ijoms.v2i9.530
[22] Suandi, F., Anam, M.K., Firdaus, M.B., et al. (2024). Enhancing sentiment analysis performance using SMOTE and majority voting in machine learning algorithms. In 7th International Conference on Applied Engineering (ICAE 2024), Atlantis Press, pp. 126-138. https://doi.org/10.2991/978-94-6463-620-8_10
[23] Rachmawanto, E.H. (2022). Sentiment analyst on twitter using the k-nearest neighbors (KNN) algorithm against covid-19 vaccination. Journal of Applied Intelligent System, 7(2): 135-145. https://doi.org/10.33633/jais.v7i2.6734
[24] Pranatawijaya, V.H., Kamala Sari, N.N., Rahman, R.A., Christian, E., Geges, S. (2024). Unveiling user sentiment: Aspect-based analysis and topic modeling of ride-hailing and google play app reviews. Journal of Information Systems Engineering & Business Intelligence, 10(3): 328-339. https://doi.org/10.20473/jisebi.10.3.328-339
[25] Anam, M.K., Putra, P.P., Malik, R.A., et al. (2025). Enhancing the performance of machine learning algorithm for intent sentiment analysis on village fund topic. Journal of Applied Data Sciences, 6(2): 1102-1115. https://doi.org/10.47738/jads.v6i2.637
[26] Talasila, V., V., M.M., R.N.M. (2023). Enhancing text-to-image synthesis with an improved semi-supervised image generation model incorporating n-gram, enhanced TF-IDF, and bow techniques. International Journal of Intelligent Systems and Applications in Engineering, 11(7s): 381-397. https://ijisae.org/index.php/IJISAE/article/view/2963.
[27] Kabra, B., Nagar, C. (2023). Convolutional neural network based sentiment analysis with TF-IDF based vectorization. Journal of Integrated Science and Technology, 11(3): 503.
[28] Omankwu, O.C.B., Osodoeke, E.C., Ubah, V.I. (2024). Disease prediction using random forest machine learning algorithm. NIPES-Journal of Science and Technology Research, 6(1). https://doi.org/10.5281/zenodo.11005308
[29] Yuliansyah, H., Imaniati, R.A.P., Wirasto, A., Wibowo, M. (2021). Predicting students graduate on time using C4. 5 algorithm. Journal of Information Systems Engineering and Business Intelligence, 7(1): 67-73. https://doi.org/10.20473/jisebi.7.1.67-73
[30] Sharupa, N.A., Rahman, M., Alvi, N., Raihan, M., Islam, A., Raihan, T. (2020). Emotion detection of twitter post

using multinomial naive bayes. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, pp. 1-6. https://doi.org/10.1109/ICCCNT49239.2020.9225432
[31] Yuliana, H., Iskandar, Hendrawan, et al. (2024). Hyperparameter optimization of random forest for 5G coverage prediction. Buletin Pos dan Telekomunikasi, 22(1): 75-90. https://doi.org/10.17933/bpostel.v22i1.390
[32] Rahayu, W., Jollyta, D., Hajjah, A., Marlim, Y.N., Desnelita, Y. (2024). Synthetic minority oversampling technique (SMOTE) for boosting the accuracy of C4.5 algorithm model. Journal of Artificial Intelligence and Engineering Applications, 3(3): 624-630. https://doi.org/10.59934/jaiea.v3i3.469
[33] Ressan, M.B., Hassan, R.F. (2022). Naïve-bayes family for sentiment analysis during COVID-19 pandemic and classification tweets. Indonesian Journal of Electrical Engineering and Computer Science, 28(1): 375-383. https://doi.org/10.11591/ijeecs.v28.i1.pp375-383
[34] Pravitasari, A.A., Asnawi, M.H., Helen, A., et al. (2024). Stravens: Sentence transformer voting ensemble for intent classification-based chatbot model. IEEE Access, 12: 197187-197200. https://doi.org/10.1109/ACCESS.2024.3519223
[35] Supriatna, D.J.I., Saputra, H., Hasan, K. (2023). Enhancing the red wine quality classification using ensemble voting classifiers. Infolitika Journal of Data Science, 1(2): 42-47. https://doi.org/10.60084/ijds.v1i2.95
[36] Karadeniz, T., Maraş, H.H., Tokdemir, G., Ergezer, H. (2023). Two majority voting classifiers applied to heart disease prediction. Applied Sciences, 13(6): 3767. https://doi.org/10.3390/app13063767
[37] Tahir, H.A., Alayed, W., Hassan, W.U., Haider, A. (2024). A novel hybrid XAI solution for autonomous vehicles: Real-time interpretability through LIME–SHAP integration. Sensors, 24(21): 6776. https://doi.org/10.3390/s24216776
[38] Salih, A.M., Raisi-Estabragh, Z., Galazzo, I.B., et al. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems, 7(1): 2400304. https://doi.org/10.1002/aisy.202400304