

Interpretable AI Pipeline for Ranking Influential Depression Factors from the Burns Depression Checklist

Tumpa Rani Shaha¹, Momotaz Begum²

^{1,2}Department of Computer Science and Engineering, Dhaka University of Engineering & Technology, Gazipur, Bangladesh

¹Department of Computer Science and Engineering, Gopalganj Science and Technology University, Gopalganj, Bangladesh

Article Info

Article history:

Received Nov 18, 2025

Revised Aug 22, 2026

Accepted Sep 9, 2026

Keywords:

Depression Screening Tool
Burns Depression Checklist
Machine Learning
Explainable Artificial-
Intelligence
Feature Selection

ABSTRACT

The Burns Depression Checklist (BDC) measures depression severity by adding the scores of 25 symptom items. However, the total score does not show which items contribute most to the severity category. This study presents an interpretable machine learning pipeline to rank predictive BDC items. We analyzed responses from 654 university students in Bangladesh. We first examined the reliability and relationships of the BDC items. We then tested seven machine learning classifiers. We also applied four feature selection methods—RFE, SKB, UFS, and MIFS—and two explanation methods, SHAP and LIME. The BDC showed strong internal consistency, with a Cronbach's alpha of 0.91. The full-item stacking model achieved 90.84% test accuracy, 86.03% balanced accuracy, and 82.60% macro-F1. We combined the six item rankings on one common scale. Loss of motivation received the highest rank. Loss of interest in work, loneliness, discouragement, hopelessness, and reduced social interest also ranked highly. These items represent reduced motivation, social withdrawal, negative mood, and negative self-view. The main contribution of this study is an item-level ranking that combines evidence from several analytical methods. However, the target categories come directly from the same 25 BDC items. Therefore, the results show how well the models reconstruct BDC severity categories, not how well they diagnose clinical depression. The rankings also do not show clinical urgency. Any positive response to a suicide related item requires separate professional assessment.

Copyright © 2026 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Momotaz Begum

Department of Computer Science and Engineering,
Dhaka University of Engineering & Technology, Gazipur, Bangladesh.

Email: drmmotaz@duet.ac.bd

1. INTRODUCTION

Depression is a complex mental health condition characterized by emotional, cognitive, behavioral, and physical symptoms of varying severity [1]. It can negatively affect relationships, education, employment, daily functioning, and physical health. It has been reported that several medical conditions, including cardiovascular and respiratory diseases, diabetes, stroke, cancer, and obesity, have a remarkable association with depression [2], [3]. It has also been associated with alcohol and substance misuse and an increased risk of suicidal behavior [4]. Depression is also an important concern in Bangladesh, particularly among students and young adults [5]. A high prevalence of depressive symptoms has been reported among different Bangladeshi populations [6]. Any psychological difficulties affect students' academic performance, social relationships, and quality of life. It has also been identified as an important factor associated with suicidal behavior among Bangladeshi students [7]. Therefore, an interpretable AI approach is needed to identify and rank the most influential depression factors. This evidence can improve understanding of symptom importance and support more focused screening and early intervention.

Over the past two decades, depression screening has received increasing attention in primary care, educational institutions, and community settings [8]. Standardized screening instruments are widely used to assess the presence and severity of depressive symptoms. These instruments support preliminary assessment, early referral, and symptom monitoring [9]. Table 1 presents ten commonly used depression screening instruments and compares their number of items, target populations, and severity categories.

Table 1. Details of the widely used ten depression screening scales

Ref	Depression Screening Scale	Symptoms considered	Applicable Participants	Degree of Depression
[10]	Burns Depression Checklist (BDC)	25	Clinical patients and Adult	No depression, Normal but unhappy, Mild, Moderate, Severe, Extreme
[11]	Beck Depression Inventory (BDI)	21	Individuals aged 13 and over	Normal, Mild, Borderline, Moderate, Severe, Extreme
[12]	Hamilton Depression Rating Scale (HDRS)	21	Hospital inpatients	Absent, Mild, Moderate, Severe and Incapacitating
[13]	Zung Self-Rating Depression Scale (ZSRDS)	20	Clinical Patients	Normal, Mild, Moderate, Severe
[14]	Geriatric Depression Scale (GDS)	15	Older patients and those suffering from dementia.	Normal, Mild, Severe
[15]	Kutcher Adolescent Depression Scale (KADS)	11	Adolescent depression	Mild, Moderate, Severe
[16]	Mongomery Asberg Depression Rating Scale (MADRS)	10	Therapeutic clinics, Patients with mood disorders	Normal, Mild, Moderate, Severe
[17]	Edinburgh Postnatal Depression Scale (EPDS)	10	Postpartum depression	Mild, Severe
[18]	Major Depression Inventory (MDI)	10	DSM-IV diagnosis	Mild, Moderate, Severe, Major
[19]	Patient Health Questionnaire (PHQ)	9	Patient Questionnaire	Minimal, Mild, Moderate, Severe

Among these instruments, the BDC is a self-reported screening instrument containing 25 symptom items. The checklist covers emotional, cognitive, interpersonal, motivational, physical, and suicidal-urge symptoms. Each item receives a numerical response, and the sum of the 25 item scores is mapped to one of six predefined severity categories: no depression, normal but unhappy, mild, moderate, severe, or extreme [20]. Although the total score provides a convenient summary of symptom severity, it does not clearly show which individual items contribute most strongly to the resulting severity.

Statistical and machine learning (ML) techniques are complementary approaches to examining questionnaire data. Correlation analysis describes the strength and direction of associations among BDC items [21]. Moreover, ML models capture complex patterns across multiple items [22]. Feature selection (FS) techniques identify the variables that contribute most strongly to a model and removes less informative variables [23]. Although ML models achieve strong predictive performance, their decisions are difficult for non-specialist users to understand. Explainable Artificial Intelligence (XAI) provides methods for examining how input variables contribute to model outputs [24]. The combination of conventional ML classifiers, FS techniques, and post hoc explainability is already common in biomedical informatics. Therefore, the methodological components of the proposed pipeline are not presented as individually novel. The specific contribution of this study is the integrated application to the BDC items and the consolidation of six methods' outputs to identify items that are consistently ranked within BDC-derived severity categories.

Accordingly, this study aims to develop an interpretable machine learning pipeline for analyzing and ranking predictive items in the BDC. The main contributions are as follows:

- Statistical reliability and correlation analyses are conducted to examine the internal consistency and relationships among the 25 BDC items.
- Seven machine learning classifiers are used to evaluate the six BDC derived severity categories.
- Four feature selection techniques (RFE, SKB, UFS, and MIFS) are compared to identify the ranking position of BDC items.
- SHAP and LIME are employed to provide global and local explanations of the selected ML model.

- The outputs of the four feature selection and two XAI methods are consolidated to identify the top ten BDC items that consistently ranked across the six analytical approaches.

The remainder of this paper is organized as follows. Section 2 reviews previous studies on depression screening, machine-learning classification, feature selection, and explainable AI. Section 3 describes the dataset, preprocessing procedure, statistical analyses, machine-learning models, feature-selection methods, XAI techniques, and consolidated ranking procedure; the complete analytical workflow is presented in Figure 1. Section 4 reports and discusses the experimental findings, compares them with existing studies, and interprets the most consistently ranked BDC items. Section 5 concludes the paper, presents its limitations, and outlines future work involving independent clinical benchmarks, external datasets, improved model validation, and safety-aware clinical evaluation.

2. LITERATURE REVIEW

Depression is common among young people and may lead to serious personal, educational, and social problems. Early screening can help identify people who may need further professional assessment. However, stigma, limited awareness, and misunderstanding of mental health conditions can prevent people from seeking help [25]. For this reason, many researchers have used depression screening questionnaires together with statistical and ML methods to support early assessment. Table 2 summarizes several studies that used ML, FS and XAI for depression assessment. It compares their participant groups, screening instruments, additional information, ML models, FS methods, and explainability techniques.

Table 2: An investigation of numerous studies conducted to predict depression.

Ref	Participants types	Screening Tools (ST)	Associated Collected Information (ACI)	Extracted the most relevant features from ACI	Applied ML model	Applied FS Algorithms	Applied XAI
[26]	Bangladeshi participants Age 16 - above	BDC	Sociodemographic, Family & personal, Financial, Physical, Habitual, Suicidal thought	✗	KNN, AdaBoost, XGBoost, Weighted Voting, Bagging	Select K Best, Boruta, mRMR	✗
[27]	Cancer Patients	BDI	Socio-demographic information, Occupational, Educational, and Economic Status	✗	ANN, Fuzzy with genetic algorithm	✗	✗
[28]	Adults with hypertension Age ≥ 40	PHQ	Socio-demographic information	✓	ANN, RF, AdaBoost, SGBost, XGBoost, and SVM	Boruta, LASSO	✗
[29]	UG students	BDI and DASS	Personal Info, Psychological Question	✗	KNN, RF, SVM	✗	✗
[30]	University students from Bangladesh Aged 19-35	Voting Algorithm	Personal, Clinical	✓	LR, GB, KNN, RF, DT, SVM, Perceptron, GNB, MNB, ZeroR, ANN, and CNN	RFE, SKB, Fisher Score, Chi-squared, ExtraTree, Pearson, MIFS, and Variance Threshold	SHAP, LIME
[31]	Geriatric patients	PHQ	Socio-demographic, psychometric information	✗	Extreme Gradient Boosting, LR	✗	✗
[32]	Mothers in their postpartum period	EPDS	Socio-economic, clinical, and psychometric information	✗	Naïve Base, LR, SVM, ANN	✗	✗
[33]	Seafarers	HDRS	Socio-demographic, occupational information	✓	CatBoost, LR, NB, RF, and SVM	RFE	✗

The BDC was used in the study [26] with data obtained from Bangladeshi participants and included sociodemographic, family, financial, physical, behavioral, and suicide related data. A number of ML and FS methods were employed, without explainable AI. In study [27] BDI was used to investigate depression among cancer patients. In study [28], the authors compared the diagnostic utility of the PHQ for predicting depression in adults with hypertension. It used several ML models and chose the proper variables with Boruta and LASSO. The results of these studies demonstrate the potential of ML to assist in analyzing depression screening data in various populations. BDI and DASS were used in study [29] to assess depression in undergraduate students. In study [30], several classifiers, FS techniques, SHAP, and LIME were used to analyze Bangladeshi university students' mental health condition. This study demonstrated the potential of applying FS with explainable AI to uncover variables affecting model outputs. However, this analysis did not focus specifically on the internal structure of the BDC with personal and clinical data. In study [31] sociodemographic and psychometric data were used to predict older adults with persistent depressive symptoms. The EPDS Scale has been used in a study [32] to evaluate postpartum depression along with some other socio-demographic, clinical, and psychological parameters. In Study [33], occupational and sociodemographic data were analysed to assess anxiety and depression in seafarers. These studies demonstrate the wide use of ML with depression screening instruments. However, it is difficult to compare the accuracy of their reported accuracy due to the different populations, input variables, definitions of target accuracy, and validation procedures used.

The literature review identifies three main gaps. First, limited research has examined the predictive importance of individual items within the 25-item BDC using FS and explainability methods together. Second, the agreement between FS results and SHAP or LIME based explanations has not been sufficiently studied for the BDC. Third, existing questionnaire based studies often give limited attention to the difference between statistical importance and clinical importance. This difference is especially important for suicide related items. An item may receive a low model based rank because it occurs less often or has limited variation, but it can still require immediate clinical attention. A previous publication by the authors [34] investigated BDC based multiclass depression severity prediction through feature grouping and ML classification. In contrast, the present study focuses on item-level statistical relationships. It then applies RFE, SKB, UFS, and MIFS to select predictive items. SHAP is used to examine the overall and class specific contribution of the items. LIME is used to explain selected individual predictions. Finally, the findings from the four feature selection and two explainability methods are merged to identify the items that are most consistently ranked within the BDC scoring framework. This distinction defines the incremental contribution of the present work.

3. RESEARCH METHOD

Figure 1 presents the complete workflow of this study. The process begins with the 25-item BDC dataset. First, the total item scores are mapped to one of the six severity categories of the BDC. The data is then prepared and examined through statistical analysis. Subsequently, the dataset is divided into training and test sets. Seven machine learning classifiers are evaluated using both the full 25-item set and four feature selection methods. The best performing model is explained using SHAP and LIME. Finally, the results of the feature selection and explanation methods are combined to identify the BDC items that are consistently ranked across the methods.

3.1. Dataset Description

This study utilized the BDC dataset previously employed in the study reported in [34]. The dataset contains responses from 654 university students in Bangladesh. The participants were between 19 and 28 years of age, and the dataset contained an equal number of male and female respondents. Each participant answered 25 BDC symptom items. These items describe emotional, cognitive, interpersonal, motivational, physical, and suicide related symptoms. Each item represented a depression symptom and was recorded as an ordered integer score from 0 to 4 (0 for "Not at all" 1 for "Somewhat", 2 for "Moderately", 3 for "A lot" and 4 for "Extremely"). A higher value indicated stronger symptom endorsement. All 25 items were used as predictors so that every ranking method assessed the complete BDC item space. BDC derived severity category was used as the target variable. The number and percentage of participants in each severity category are reported in Table 3 to make the class distribution clear.

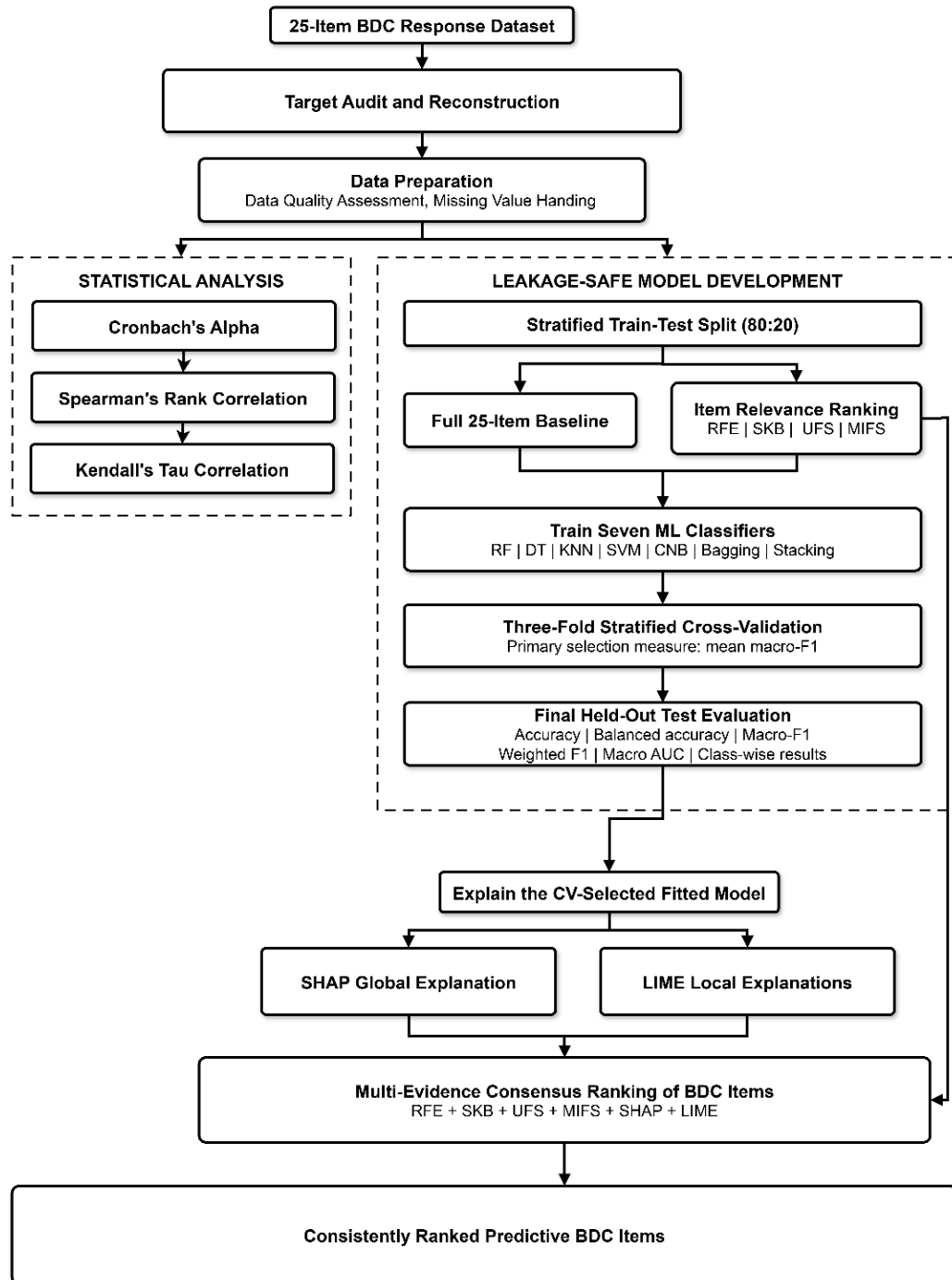


Figure 1. Proposed interpretable machine-learning pipeline for reconstructing BDC-derived severity categories and ranking predictive BDC items

Table 3 presents the distribution of the 654 participants across the six BDC derived severity categories according to score range. Most participants were classified in the “Moderate” category (55.05%), followed by “Mild” (27.52%) and “Severe” (10.70%). The “No depression”, “Normal but Unhappy”, and “Extreme” categories contained relatively few participants.

The dataset did not include a diagnosis made by a psychiatrist or another mental health professional. It also did not include a Structured Clinical Interview for DSM disorders, a DSM-5-based interview, or another independent clinical reference standard. Therefore, the outcome represents self-reported BDC severity and is not interpreted as a confirmed clinical diagnosis.

Table 3. Distribution of BDC-derived severity categories

Severity category	Score range	Participants	Percentage
No depression	0–5	14	2.14%
Normal but unhappy	6–10	23	3.52%
Mild	11–25	180	27.52%
Moderate	26–50	360	55.05%
Severe	51–75	70	10.70%
Extreme	76–100	7	1.07%
Total	—	654	100.00%

3.2. Target Reconstruction and Analytical Scope

For each participant BDC total score was recalculated by adding the scores of all 25 items. Then total score was mapped to the six score ranges reported in Table 3. The stored dataset labels were audited against the reconstructed total scores. The dataset rule produced zero mismatches. This creates a deterministic and circular relationship between the predictors and the target. The machine-learning task therefore measures how well the models reconstruct the BDC scoring rule. It does not measure how well the models predict an independently observed clinical diagnosis. The conventional BDC total score calculation was treated as the internal rule based reference. When all 25 scores are available, this rule reproduces the severity category exactly by definition. Machine learning was retained to examine whether some items contain more information about the score bands than others and whether their rankings remain similar across different methods. All reported model scores are therefore described as reconstruction performance.

3.3. Data Preparation

The dataset was checked for its dimensions, column names, missing values, invalid values, response ranges, and duplicate rows. All predictors were already stored as ordered integer scores, so categorical encoding was not required. Any learned preprocessing step was placed inside the modelling pipeline and fitted only from the relevant training data.

3.4. Statistical Analysis

We examined the internal structure of the BDC responses by applying three statistical methods: 1. Cronbach's alpha [35], 2. Spearman's rank correlation [36], and 3. Kendall's tau-b correlation [37].

3.4.1. Internal Consistency

At this stage, we use 'Cronbach's alpha' to measure the internal consistency of the 25 items of the BDC. This measure indicates whether the items relate to the same broader construct. A high alpha value reflects strong internal consistency within the current sample [35]. However, this value alone does not establish diagnostic validity, construct validity, or the generalizability of the findings to other populations.

3.4.2. Spearman's Rank Correlation

Spearman's rank correlation was used to measure monotonic relationships among the ordered BDC item scores. A positive coefficient indicates that two variables generally increase together. A negative coefficient indicates that one item bends to decrease as the other increases. Values near zero indicate a weak monotonic relationship [36].

3.4.3. Kendall's Tau-b Correlation

Kendall's tau-b was also used because it is suitable for ordered data and accounts for tied values [37]. Spearman's rho and Kendall's tau-b provide related but different views of association. The correlations were used to describe relationships among BDC items. They do not provide evidence of causality or independent clinical validity.

3.5. Leakage-Safe Data Partitioning

The prepared dataset was divided into training and test sets using a stratified 80:20 split and a random seed of 42. It preserved the proportion of the six severity categories as closely as possible. The test set was isolated before model fitting, or model selection. Three-fold stratified cross-validation was performed within the training set. Feature selection were fitted separately inside each training fold.

3.6. Full-Item Baseline and Item-Relevance Assessment

The training analysis included a full 25-item baseline and four item-relevance methods: recursive feature elimination (RFE), SelectKBest with chi-square (SKB-Chi²), univariate ANOVA F-test (UFS), and

mutual-information feature selection (MIFS). These methods generated comparative rankings. They were not used to claim that a reduced questionnaire outperformed the original scale.

3.6.1. Recursive Feature Elimination

RFE repeatedly fitted an estimator and removed the least important remaining item. The elimination order was converted into an item rank. RFE eliminate items to produce a genuine ordering, although the final predictive model retain all 25 items. If all 25 items are retained inside RFE, no elimination occurs and the ranks are tied [38].

3.6.2. SelectKBest with Chi-Square

SKB- χ^2 measured dependence between each non-negative item and the severity category. A larger chi-square statistic indicated stronger dependence and produced a higher rank [39].

3.6.3. Univariate ANOVA F-Test

UFS measured between class variation relative to within class variation for each item. A larger F-value indicate stronger separation among the six severity categories [40].

3.6.4. Mutual-Information Feature Selection

MIFS measured the information shared between an item and the target, including possible nonlinear dependence. Items with larger mutual information values received higher ranks [41].

3.7. Machine-Learning Classifiers

Seven supervised ML classifiers were evaluated: Random Forest (RF), Decision Tree (DT), K-nearest Neighbors (KNN), Support Vector Machine (SVM), Categorical Naive Bayes (CNB), Bagging, and Stacking [22].

3.7.1. Decision Tree

A Decision Tree divides the data into smaller groups by applying a sequence of decision rules. The model can handle multiclass outcomes and is easy to interpret. However, a deep tree may fit the training data too closely and perform poorly on new observations [42].

3.7.2. Random Forest

RF combines several decision trees. Each tree is trained using a sample of the training data and a subset of the input variables. The final class is selected by combining the predictions of the trees. This process can reduce the instability of a single decision tree [43].

3.7.3. K-Nearest Neighbors

KNN classifies a new observation using the classes of its nearest training observations [44]. The value of (k) determines how many neighboring observations are considered. Because KNN depends on distance, the input variables placed on a comparable scale.

3.7.4. Support Vector Machine

SVM identifies boundaries that separate the classes. It can model both linear and nonlinear relationships through the use of different kernel functions. The penalty parameter and kernel settings selected using the training data [45].

3.7.5. Categorical Naive Bayes

CNB is a probabilistic classifier that was developed to improve classification when the class distribution is uneven. It estimates the probability of each class from the observed input values. Its suitability for the BDC response scale explained and checked during model evaluation [46].

3.7.6. Bagging

The Bagging classifier was used as an ensemble model to improve prediction stability. A Multilayer Perceptron (MLP) neural network was used as the base classifier. The MLP contained one hidden layer with 100 neurons and was trained for a maximum of 1,000 iterations. Early stopping was enabled to reduce unnecessary training and control overfitting. The Bagging model combined 10 MLP classifiers. Each classifier was trained on a different bootstrap sample of the training data, and their predictions were combined to produce

the final prediction. Before model training, the input features were standardized using StandardScaler because MLP performance affected by differences in feature scale [47].

3.7.7. Stacking

Stacking classifier was used to combine the prediction ability of different machine learning models [48]. Four classifiers were used as base learners: DT, RF, CNB, KNN. These models were selected to combine different learning approaches within one ensemble framework. The DT used to balanced the class weights. The RF contained 200 trees and used to balanced subsample class weighting. CNB was applied with an alpha value of 1.0. The KNN classifier used five nearest neighbors with distance based weighting. StandardScaler was applied within the KNN pipeline to standardize the input features. The predicted class probabilities from the four base classifiers were passed to a RF meta classifier. The meta classifier contained 200 trees and used balanced subsample class weighting. Three fold internal cross-validation was used to generate the predictions for training the meta classifier. Therefore, the final prediction was not based on simple majority voting. The RF meta classifier combine the probability outputs produced by the four base classifiers. A random state of 42 was used where supported, and parallel processing was enabled using all available CPU cores. This stacking structure allowed different types of classifiers to contribute to the final depression severity prediction.

3.8. Classifier Development and Cross-Validation

Total 35 combinations of five feature settings and seven classifiers were evaluated through three fold stratified cross-validation on the training set. Macro-F1 was used as the primary model selection measurement. It gives equal importance to each class and is less influenced by the large moderate severity group.

3.9. Final Held-Out Performance Evaluation

The selected pipeline was refitted on the complete training set and evaluated once on the held-out test set. Accuracy, balanced accuracy, macro precision, macro sensitivity, macro specificity, macro-F1, weighted F1, and macro one-versus-rest ROC-AUC were calculated. Class-wise results and the confusion matrix were also examined because the no depression and extreme categories contained very few observations.

3.10. Explanation of the Selected Model

SHAP and LIME were applied after model selection. These methods explain how the selected model uses the BDC items. They do not identify the clinical causes of depression.

3.10.1. SHAP Global Explanation

Model-agnostic permutation SHAP estimated global item contribution across held out observations. Mean absolute SHAP values were used to measure contribution strength, and items were ranked in descending order [49].

3.10.2. LIME Local Explanations

LIME fitted simple surrogate models around reproducibly selected test cases. One available case per severity category was explained, and the absolute local weights were aggregated. LIME was treated as supporting local evidence rather than a complete global profile [50].

3.11. Multi-Evidence Consensus Ranking of BDC Items

A multi-evidence consensus approach was used to combine the item rankings produced by six methods. These methods included RFE, SKB, UFS, MIFS, SHAP, and LIME. The first four methods provided statistical feature selection evidence, while SHAP and LIME provided model explanation evidence. The combination of these methods has provided the opportunity to evaluate each BDC item from various analytical perspectives. Before consolidation, the method labels were checked to confirm that the ranking results contained all six expected methods. The ranking tables were then combined into a single table. For each method (m), an item (j) received a binary vote according to its selection status:

$$v_{jm} = \begin{cases} 1, & \text{if item } j \text{ was selected by method } m, \\ 0, & \text{otherwise.} \end{cases}$$

When selection was defined using the prespecified list, the same rule can be expressed as:

$$v_{jm} = \begin{cases} 1, & r_{jm} \leq 25, \\ 0, & r_{jm} > 25, \end{cases}$$

where (r_{jm}) represents the rank of item (j) under method (m).

The total occurrence count for each item was determined by summing the votes obtained through six methods:

$$O_j = \sum_{m=1}^6 v_{jm}$$

The range of occurrence counts was between 0 and 6. A higher value indicated that the item in question had been selected in a greater number of ways, thereby reflecting a strong consensus across the six sources of evidence. Since multiple items could share the same occurrence count, additional rank statistics were calculated. The average rank for item (j) was determined as follows:

$$\bar{r}_j = \frac{1}{6} \sum_{m=1}^6 r_{jm}$$

Smaller mean ranks meant that the item, on the whole, was given a higher rank by the six methods. The median rank was also computed as the middle ranking position. In addition, the highest and lowest positions given to items were recorded in the best rank and the worst rank. These measures were employed to illustrate the agreement between the methods for the item rankings. A hierarchical sorting rule was used to determine the final consensus ranking. Items were initially listed by the number of methods in which they occurred, with the most frequently occurring items listed first. If two or more items had the same number of occurrences, the one with the lower mean rank was given a higher rank. In the event of a second tie, the next lower rank of the smaller tie was used. The name of the item was used alphabetically as the last tie breaker. The data was sorted and then the data was assigned sequential consolidated ranks from 1 to 25. The final list of consistently ranked predictive BDC items was made up of the first 10 of the items. Six-method rankings were saved for all of the methods, consolidated rankings for all items were saved and top-10 list was saved for further analysis. Three BDC items were singled out regarding suicide and received a separate clinical-safety flag: suicidal thoughts, desire to end one's life, and plan for self-harm. The statistical consensus rank was not affected by this flag. It was added to illustrate that the concepts of statistical and clinical importance are different. So if a person is responding positively to a suicide item, regardless of whether that item is in the top or bottom spot in the predictive ranking, it should be assessed by the appropriate professional.

3.12. Clinical Safety and Interpretation Boundary

Consolidated ranking should not be used as a clinical triage hierarchy. A symptom with a lower clinical rank can be an important symptom despite being a lower ranked symptom from a statistical point of view. This is particularly true for the three suicide related BDC items. No matter how many times suicidal ideation, self-destructive thoughts, or self-destructive plans appear, or how many points are assigned to them according to the SHAP range, or how many points are given for their appearance in the LIME or how they rank on the final list, suicidal thoughts, suicidal ideation, and suicidal plans with regard to self-destructive behaviour must be assessed professionally. The purpose of the pipeline is to conduct exploratory studies of BDC item patterns, and not to used as a substitute for assessment by a skilled mental health worker.

4. RESULTS AND DISCUSSION

In this section, the results of the proposed approach are presented in the following workflow order: target auditing, statistical evaluation, model selection based on training only, held-out evaluation, item ranking, model interpretation, consensus ranking, and clinical interpretation.

4.1. Target Reconstruction and Audit of the Severity

Each of the 25 item responses was scored and summed to make a total BDC score for each observation. The total score was then divided into the following prespecified severity ranges: no depression (0-5), normal but unhappy (6-10), mild depression (11-25), moderate depression (26-50), severe depression (51-75) and extreme depression (76-100). The reconstructed labels were matched with the severity labels that were already in the data set. Equivalent observed 30 times, which equals 4.59% of the observations. The mismatches were at boundary scores (5, 10, 25, 50 and 75). This pattern suggests that the interval boundaries were most likely stored with a different rule. Therefore, by definition, the BDC scoring rule is an internal reference (IR) that is designed to recreate the category with 100% agreement. The machine-learning results indicate the performance of the various models on approximating this rule, and identify the items that are most stable as providing predictive evidence within the BDC framework.

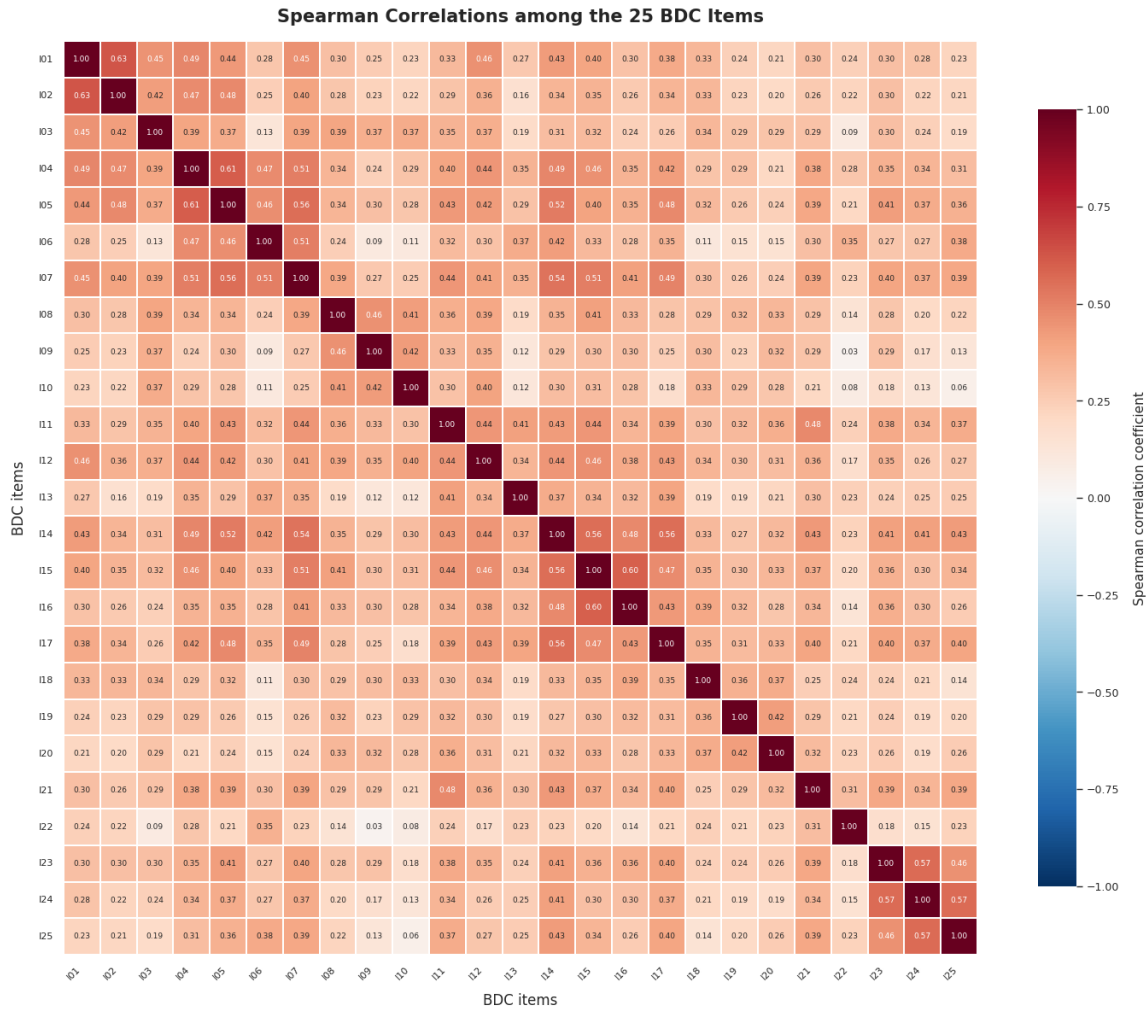


Figure 2. Spearman correlation heatmap for the 25 BDC items.

4.2. Reliability and Inter-Item Correlation

The 25 BDC items had a Cronbach's alpha of 0.91. This score reflects good internal consistency of the current sample. It implies that the items tend to assess related aspects of depression symptom severity. It does not, however, make a clinical diagnosis or confirm it to be valid in other populations. Spearman correlation and Kendall's tau-b correlation was also computed for all item pairs. The above methods were appropriate due to the ordinal nature of the BDC responses and tied values. The heatmaps revealed the degree of association of items and enabled the identification of areas of common information.

Results of the Spearman and Kendall's tau-b analyses indicated that there was a positive correlation between all 25 BDCs as indicated in Figure 2 and Figure 3. Spearman's correlation coefficients were about 0.03 to 0.63 and Kendall's tau-b coefficients were 0.03 to 0.56. The strength of most relationships was weak to moderate. This discovery suggests that the items assess different dimensions of depression, but there is not the same information. Spearman's ($\rho=0.63$) and Kendall's ($\tau_b=0.56$) correlation coefficients were the highest ones for I01 with I02. Other relatively strong relationships appeared between I04 and I05 ($\rho=0.61$), ($\tau_b=0.55$), I15 and I16 ($\rho=0.60$), ($\tau_b=0.54$), I23 and I24 ($\rho=0.57$), ($\tau_b=0.54$), and I24 and I25 ($\rho=0.57$), ($\tau_b=0.55$). The consistency of the observed relationships was confirmed by similar patterns in both measures. There was no excessively high correlation for any pair. The results did not therefore show the items as completely redundant in the BDC. Rather, the shared and unique content led to the decision to keep all items for individual relevance assessment. It also enabled the use of the proposed interpretable AI pipeline for comparing the importance of items using various feature-selection techniques and explanations. For the correlation analysis, relationships among the items were described while for the proposed consensus ranking, the relative predictive relevance to BDC-derived depression severity was identified.

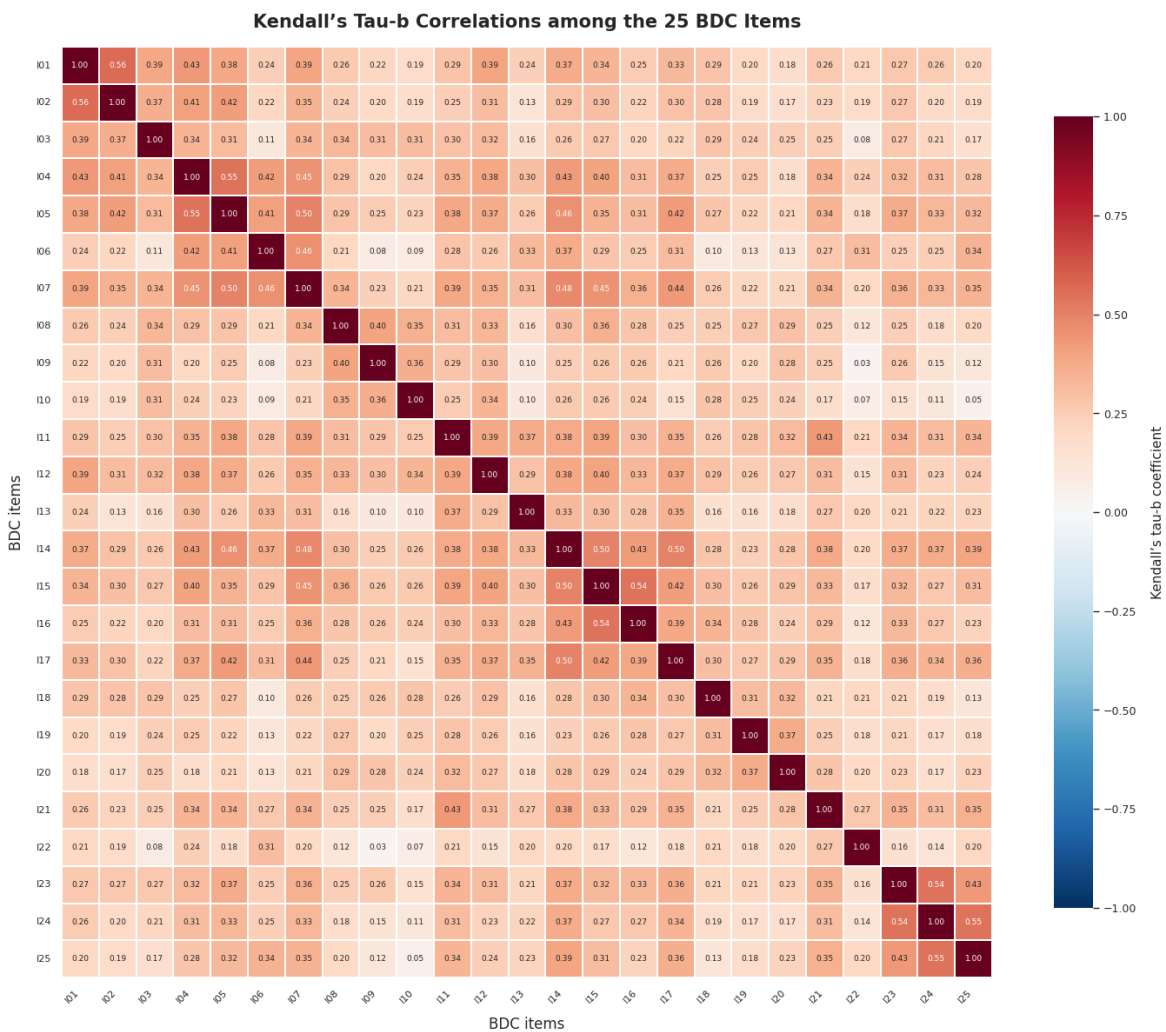


Figure 3. Kendall's tau-b correlation heatmap for the 25 BDC items.

4.3. Training Cross-Validation and Model Selection

Five item-set configurations and seven classifiers produced 35 fixed modelling configurations. All 35 feature-setting and classifier combinations were compared through three-fold stratified cross-validation using training data only. Table 4 presents the strongest configuration under each feature setting.

Table 4. Best training cross-validation result under each feature setting

Feature setting	Best classifier	Mean macro-F1	Standard deviation
Full 25-item set	Stacking	0.798	0.155
RFE	SVM	0.713	0.072
SKB-chi-square	Stacking	0.700	0.014
UFS-F score	Stacking	0.782	0.114
MIFS	Stacking	0.759	0.057

Macro-F1 was defined before evaluation as the primary selection measure. This choice was appropriate because macro-F1 gives the same weight to every severity class and is less influenced by the majority category than accuracy. The full 25-item Stacking model achieved the highest mean macro-F1 of 0.798 and was selected before the test set was examined. UFS-Stacking and MIFS-Stacking produced the next highest results. None of the four selected 25-item subsets exceeded the full-item Stacking model in training cross-validation. Feature selection remained useful for ranking analysis, but it did not improve the primary model selection. The relatively large standard deviation of the selected model reproduces dissimilarity across the three folds. The “No Depression” and “Extreme” groups paid to this indecision. Repeated stratified validation and a larger external sample would provide a more stable estimate

4.4. Held-Out Performance of the Selected Stacking Model

Once the model selection was completed, the full pipeline of stacking with all items was fitted to the training set and tested once with the held-out test set. Table 5 shows the accuracy of 0.885 and weighted-F1 of 0.880 when using the model. These results indicate that the majority of observations, particularly in the larger categories, were correctly reconstructed.

Table 5. Overall held-out performance of the selected stacking model

Performance measure	Value	Meaning in this analysis
Accuracy	0.885	Overall proportion of correct category predictions
Balanced accuracy	0.675	Average recall across severity categories
Macro precision	0.637	Average class-specific precision
Macro sensitivity	0.675	Average class-specific recall
Macro specificity	0.974	Average one-vs-rest specificity
Macro-F1	0.649	Equal-weight average of class F1-scores
Weighted-F1	0.880	F1 weighted by category size
Macro ROC-AUC (OvR)	0.990	Average probability-ranking ability

From the analysis, we found that the balanced accuracy and macro-F1 was 0.675 and 0.649 respectively. These values were lower than the overall accuracy. This gap indicates that the strong result was not shared equally by all categories. Macro specificity was high at 0.974; it means that the model avoided assigning observations to incorrect one-versus-rest categories. However, this value should be read together with sensitivity because high specificity does not guarantee that rare positive cases are detected

4.5. Class-Wise Performance and Error Pattern

In Table 6, the “Moderate depression” produced the highest F1-score of 0.941. The “Mild depression” showed strong performance, with an F1-score of 0.897. These were the two largest categories in the test set. The severe category produced an F1-score of 0.800.

Table 6. Class-wise held-out performance

Severity category	Precision	Recall	F1-score
No depression	0.750	1.000	0.857
Normal but unhappy	0.500	0.333	0.400
Mild depression	0.854	0.946	0.897
Moderate depression	0.970	0.914	0.941
Severe depression	0.750	0.857	0.800
Extreme depression	0.000	0.000	0.000

Performance was weaker for the smaller categories. Recall for the normal-but-unhappy category was 0.333, meaning that only two of its six observations were correctly identified. The single extreme observation was not correctly classified. Although recall for the no-depression category was 1.000, this estimate was based on only three observations and should not be treated as a stable estimate of general performance.

The class-wise results explain the gap between accuracy and macro-F1. Mild and moderate depression together represented 107 of the 131 test observations, or 81.68%. Good performance in these large groups strongly influenced accuracy and weighted-F1. Macro-F1 gave the rare categories equal weight and therefore exposed the weaker results in those groups.

4.6. Interpretation of the High ROC-AUC

The macro one-versus-rest ROC-AUC was 0.990. This value was much higher than the macro-F1 of 0.649. The two measures evaluate different properties. ROC-AUC measures whether predicted probabilities rank positive observations above negative observations across many possible thresholds. In contrast, F1 is based on the final predicted class after one decision rule is applied. A model can therefore produce a high AUC while still assigning an incorrect final label to a rare observation. This occurred in the extreme category, where the only test observation was missed despite the high overall macro-AUC. The AUC result indicates strong probability separation in this sample, but it should not override the confusion matrix or class-wise recall. Its uncertainty is also high because some classes had very few observations.

4.7. Descriptive Test-Set Comparison of All Configurations

A descriptive test-set comparison of all 35 fixed configurations was performed. The SVM achieved a test accuracy of 0.908, which was higher than the selected stacking model’s accuracy of 0.885. This result does not justify changing the selected model because stacking had already been chosen using training-only cross-validation and macro-F1. Selecting SVM after examining the held-out results would use the test set for model selection and could give an optimistic estimate. The primary test result should remain the result of the cross-validation-selected stacking model.

4.8. Ranking by Four Item-Relevance Methods

RFE, SelectKBest, UFS, and MIFS were fitted using only the training data. This training-only fitting reduced information leakage between model development and final testing. Each method examined item relevance from a different statistical view. RFE used multivariable model contribution, chi-square measured dependence between item values and severity, UFS measured between-category separation, and mutual information captured both linear and nonlinear dependence. Because all 25 items were retained, these procedures generated full rankings rather than reduced feature subsets. The results show which items were more informative under each statistical criterion. They do not show that lower-ranked items are clinically unnecessary.

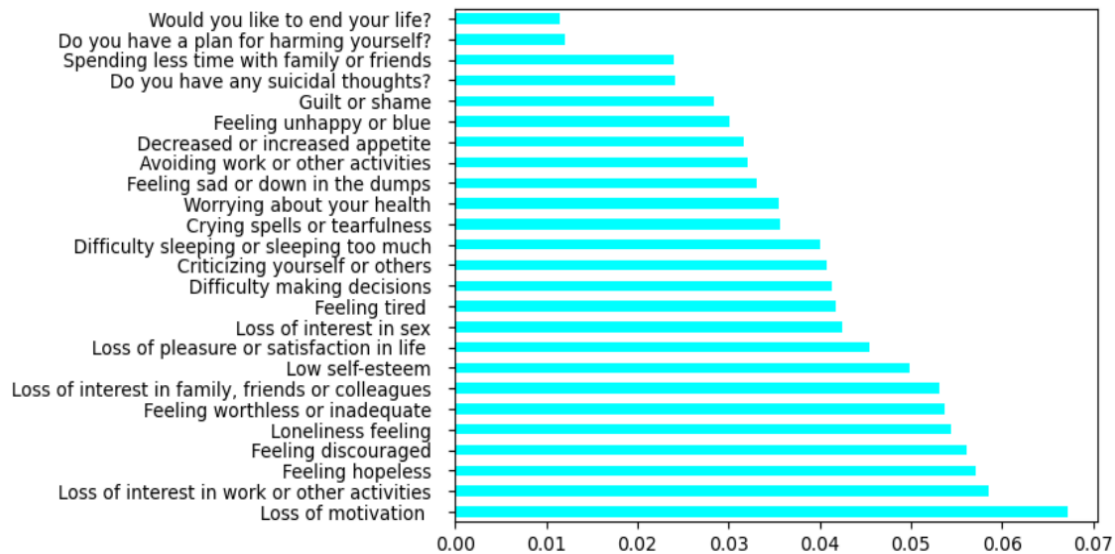


Figure 4. Feature importance ranking of the 25 BDC items using RFE.

In Figure 4, RFE analysis ranked loss of motivation as the most important item, followed by loss of interest in work or other activities and feeling hopeless. This shows that motivational and emotional symptoms contributed strongly to BDC severity prediction. These rankings indicate predictive importance, not clinical priority or causal influence.

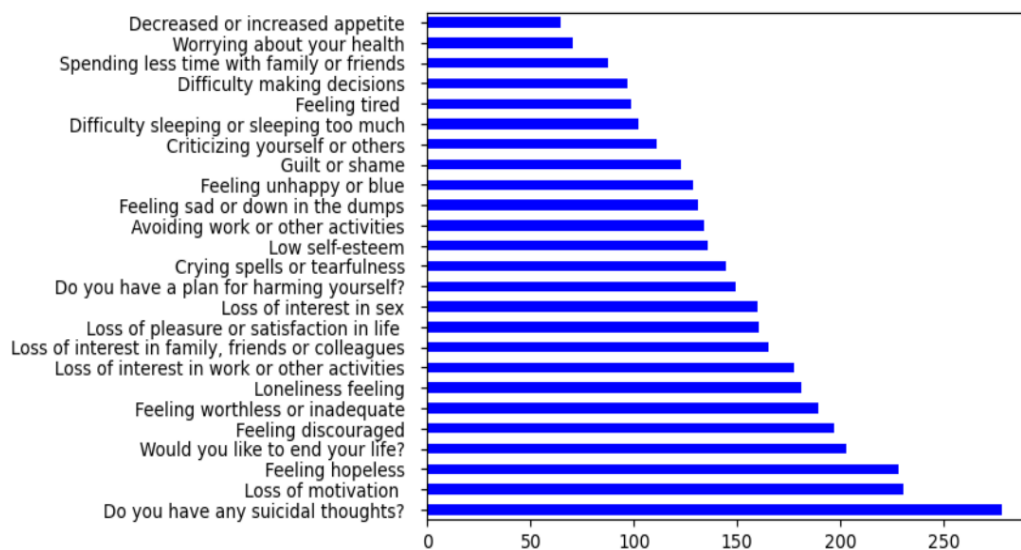


Figure 5. Feature importance ranking of the 25 BDC items using the SKB.

In Figure 5, SKB ranked suicidal thoughts as the most important item. Loss of motivation and feeling hopeless were the next highest-ranked items. These items showed the strongest statistical relationship with BDC severity.

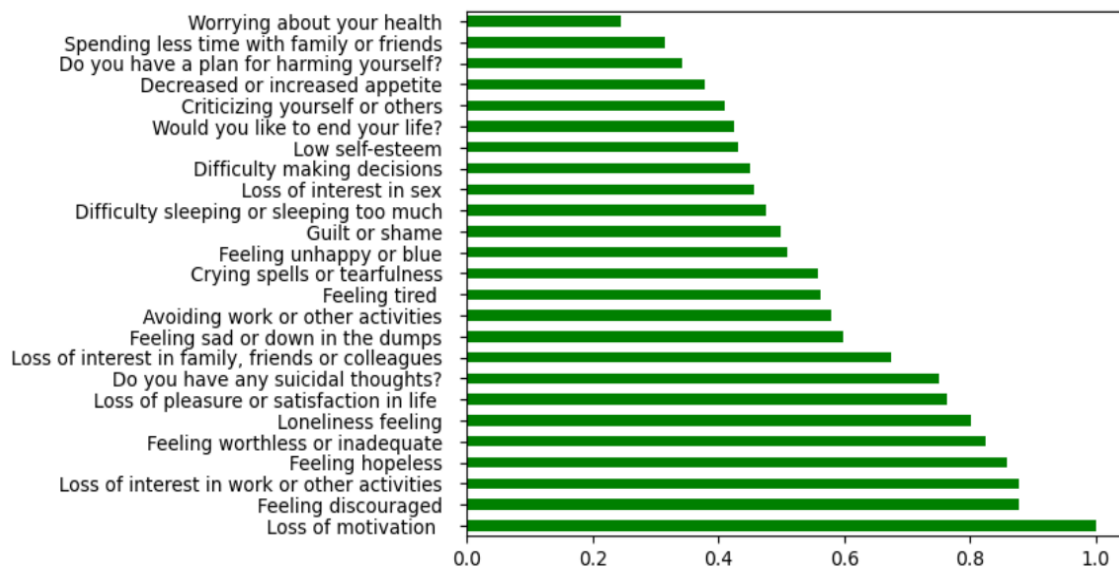


Figure 6. Feature importance ranking of the 25 BDC items using UFS.

In Figure 6, univariate feature selection analysis ranked loss of motivation as the most important item. Feeling discouraged and loss of interest in work or other activities were the next highest-ranked items. These items showed strong individual relationships with BDC severity.

In Figure 7, the mutual information analysis ranked loss of pleasure or satisfaction in life as the most important item. Loss of interest in family, friends or colleagues and loss of motivation were the next highest-ranked items. These items showed strong relationships with BDC severity.



Figure 7. Feature importance ranking of the 25 BDC items using MIFS

4.9. SHAP Global Explanation

Model-agnostic permutation SHAP was applied to the cross-validation-selected stacking pipeline. SHAP values were calculated for a deterministic sample of 100 held-out observations and were averaged in absolute form across observations and classes. The resulting values described how strongly each item changed the model's predicted probabilities, regardless of direction, are given in Table 7.

Table 7. Ten highest-ranked items according to SHAP

SHAP rank	BDC item	Mean absolute SHAP value
1	Loneliness feeling	0.016856
2	Loss of motivation	0.016797
3	Loss of interest in family, friends or colleagues	0.016727
4	Loss of interest in work or other activities	0.015756
5	Low self-esteem	0.014812
6	Loss of interest in sex	0.014526
7	Feeling hopeless	0.013675
8	Feeling discouraged	0.013524
9	Crying spells or tearfulness	0.013356
10	Difficulty making decisions	0.012192

Loneliness was the highest-ranked SHAP item, followed closely by loss of motivation and loss of interest in family, friends, or colleagues. This pattern indicates that social disconnection and reduced motivation had strong influence on the probability outputs of the selected model. Because absolute SHAP values were used, the table reports the size of influence but not whether a higher item score increased or decreased a particular class probability.

4.10. LIME Local Explanation

LIME was used to explain one deterministic test observation from each available severity class. The absolute local weights were then averaged across these representative cases. Loss of interest in sex was ranked first, followed by having a plan for self-harm, low self-esteem, feeling sad, and crying spells or tearfulness according to Table 8.

Table 8. Ten highest-ranked items in the exploratory LIME aggregation

LIME rank	BDC item	Mean absolute LIME weight
1	Loss of interest in sex	0.050262
2	Do you have a plan for harming yourself?	0.041132
3	Low self-esteem	0.040649
4	Feeling sad or down in the dumps	0.037489
5	Crying spells or tearfulness	0.037470
6	Loss of pleasure or satisfaction in life	0.037420
7	Loss of interest in family, friends or colleagues	0.034401
8	Feeling tired	0.034149
9	Loneliness feeling	0.034139
10	Loss of motivation	0.032031

These results show how selected individual predictions were formed. However, only one case per class was used. The averaged LIME result should therefore be described as an exploratory summary of representative local explanations, not as a stable population-level importance estimate. The difference between the SHAP and LIME leaders is expected because SHAP summarized a larger sample, whereas LIME focused on a small set of individual predictions.

4.11. Evaluation of Multi-Evidence Consensus Ranking of BDC Items

The rankings from RFE, SKB, UFS, MIFS, SHAP, and LIME were brought together to obtain one transparent summary. The final order used occurrence count first and mean rank as a tie-breaker. However, the current setting marked all 25 items as selected by every method because the selection threshold was 25. Occurrence count did not separate the items, and the final order was determined mainly by the mean rank. For this reason, the result is most accurately described as a multi-evidence rank aggregation. It is not an approach to feature subset reduction. The contribution of this step is that it compares several forms of evidence on one common rank scale. This reduces dependence on a single ranking method, but it does not prove that the methods are statistically independent or clinically equivalent. According to the findings, loss of motivation achieved the strongest average position, with a mean rank of 2.83. It was followed by loss of interest in work or other activities, loneliness, feeling discouraged, and feeling hopeless. The leading set covered motivational, social, affective, and pleasure-related symptoms. This broad coverage recommends that the model did not rely on only one symptom domain.

Table 9. Ten leading items in the consolidated mean rank order

Final rank	BDC item	Mean rank
1	Loss of motivation	2.83
2	Loss of interest in work or other activities	4.50
3	Loneliness feeling	5.00
4	Feeling discouraged	5.83
5	Feeling hopeless	6.00
6	Loss of interest in family, friends or colleagues	6.50
7	Loss of pleasure or satisfaction in life	7.67
8	Feeling sad or down in the dumps	7.83
9	Loss of interest in sex	8.17
10	Low self-esteem	8.50

Table 9 lists the ten highest ranked BDC items. The consolidated features cover major clinically recognizable symptom domains and common signs of depression. Loss of motivation, interest in work, pleasure, and interest in sex show reduced interest and activity. Loneliness and loss of interest in family or friends show social withdrawal. Discouragement, hopelessness, sadness, and low self-esteem reflect negative mood and self-view. The six methods ranked these items highly because they provided useful information for reconstructing the BDC severity categories. Earlier questionnaire based studies also identified mood, motivation, loss of interest, and social withdrawal as useful predictors of depression. However, this study compares individual BDC items across four FS methods and two XAI methods using one common ranking scale. These results show importance within the BDC scoring system, but they do not prove clinical validity or causality. Suicide-related items are not in the top ten, but any positive response to these items still requires immediate professional assessment.

4.12. Clinical-Safety Interpretation

Three suicide-related questions were distinguished from the statistical ranking. This separation was necessary because predictive rank and clinical urgency are not the same thing. A suicide-related item might have a lower average predictive rank simply because it occurs less frequently; nevertheless, a positive response to such an item necessitates immediate professional assessment in accordance with appropriate clinical protocols. In the exploratory LIME ranking, the item regarding ‘plans for self-harm’ emerged in second place, indicating that it exerted a strong influence on certain individual predictions. This result should not be used to compare the urgency or severity of different suicide-related responses. Statistical ranking can never serve as a substitute for established methods of risk assessment and referral.

4.13. Comparison with Related Studies

The present findings extend the depression-prediction studies summarized in Table 2 in two ways. Prior work commonly combined a depression scale with sociodemographic, clinical, educational, occupational, or lifestyle variables [26]-[33]. Some studies applied feature selection [26], [29], [30], [33] and one incorporated both SHAP and LIME [30]. Direct comparison using accuracy alone would therefore be misleading because the studies used different populations, predictors, target definitions, class distributions, and validation procedures. In contrast, the present study focuses specifically on the internal symptom items of the BDC and consolidates multi evidence rankings from four feature-selection methods and two XAI methods to obtain a cross-method list of consistently ranked questionnaire features. Additionally, the proposed approach adds a target reconstruction audit, training fold feature selection, a full item baseline and macro-averaged evaluation. The contribution is not a new classifier or a new XAI algorithm. It is a systematic item level analysis of the 25-item BDC and a transparent comparison of ranking evidence across different methods. Under the leakage safe procedure, the full-item Stacking model was selected through training only cross validation. This difference shows that data splitting, leakage control, and the model selection measure can materially affect the reported conclusion. Therefore, the main novelty is the BDC-specific, multi-method consolidation of symptom-level evidence rather than a claim of universal predictive superiority.

4.14. Practical Implications and Future Model Improvements

In clinical or counseling settings, the ranked BDC items can serve as decision-support indicators, helping professionals prioritize follow-up questions. Furthermore, the study findings help identify which BDC items provide the most information regarding different levels of severity. Item-specific interpretations can facilitate a more precise analysis of symptoms such as loss of motivation, hopelessness, reduced pleasure, and social withdrawal. However, the proposed method is not yet ready to independently determine clinical diagnoses or treatment decisions. A qualified professional must review the complete questionnaire alongside other clinical data. Statistical ranking must be distinguished from clinical urgency; specifically, a positive response regarding suicidality requires immediate assessment, even if the item’s predictive rank is low.

Several improvements are needed before clinical implementation. Future research should test the model on larger, more diverse samples and compare the results with clinician diagnoses, structured interviews, or other established screening tools. Repeated or nested cross-validation and external validation could provide a more reliable assessment of the model's performance. 'Class-balancing' techniques could be employed to improve results for rare severity levels, though these should be applied only within the 'training folds.' Additionally, future studies should examine probability calibration, ranking stability, fairness across participant groups, and sensitivity to suicide-related responses. These measures could enhance the model's reliability, safety, and practical utility.

5. CONCLUSION

The findings show that the relevance of a BDC item depends on how it is examined. Some methods are based on statistical relationships, others are based on predictive contribution, or on model behaviour. The proposed Multi Evidence consensus ranking of BDC items bundles the various types of evidence. It is useful for determining what is applicable across multiple analysis perspectives and not solely based on a single perspective. The findings were extremely consistent for items of loss of motivation, decreased pleasurable activities, hopelessness, worthlessness, discouragement and social withdrawal. The results offer an item-level view which is not obtainable from the overall BDC score. They might assist the researchers to understand which points of the checklist have more information in its scoring system. The study is limited in the study's construction. The item responses provided for the questions were directly used to derive the severities. The observed performance can thus only be interpreted as a reinstatement of the BDC scoring structure, and cannot be used to establish independent diagnostic validity. Limited generalization because of use of one student data set. Moreover, less common severity categories can result in volatile class-specific results. Special attention needs to be paid to items related to suicide. They cannot be demeaned by their statistical position and do not become "statistically insignificant" because of it. Responses to the items with lower consensus rank that mention suicidal thoughts, suicidal intent, or suicidal plan should result in an appropriate safety evaluation. Further studies should be conducted to see if the item patterns found do hold for larger, more diverse and external samples. Independent that is clinically evaluated results are required to confirm the value of the diagnosis. These would be stronger forms of evidence: rank-stability analysis and repeated or nested validation, probability calibration, fairness assessment, and explicit evaluation of sensitivity to suicide-risk. Until such validation is available, the framework should be regarded as an exploratory tool for understanding item-level predictive agreement within the BDC.

ACKNOWLEDGMENTS

This work was supported by the Information and Communication Technology (ICT) Division, under the Ministry of Posts, Telecommunications, and Information Technology, The People's Republic of Bangladesh.

CONFLICT OF INTERESTS

The authors have no competing interests to declare that are relevant to the content of this article.

ETHICS APPROVAL

Not applicable.

CONSENT TO PARTICIPATE

Not applicable.

CONSENT FOR PUBLICATION

Not applicable.

DATA AVAILABILITY

The data that support the findings of this study can be obtained from the corresponding authors upon reasonable request.

AUTHORS' CONTRIBUTIONS

Tumpa Rani Shaha: Conceptualization, Methodology, Data Curation, Software, Formal Analysis, Visualization, Writing – Original Draft. Momotaz Begum: Supervision, Validation, Review & Editing, Project

REFERENCES

- [1] G. Dunn, P. Sham, and D. Hand, "Statistics and the nature of depression," *Psychol. Med.*, vol. 23, no. 4, pp. 871–889, Nov. 1993, doi: 10.1017/S0033291700026350.
- [2] M. Mohit, M. Maruf, H. Ahmed, and M. Alam, "Depression and Physical Illnesses: an Update," *Bangladesh Med. J.*, vol. 40, no. 1, pp. 53–58, Jan. 1970, doi: 10.3329/bmj.v40i1.9966.
- [3] X. Li, J. Zhou, M. Wang, C. Yang, and G. Sun, "Cardiovascular disease and depression: a narrative review," *Front. Cardiovasc. Med.*, vol. 10, p. 1274595, Nov. 2023, doi: 10.3389/fcvm.2023.1274595.
- [4] N. Sakib *et al.*, "Depression and suicidality among Bangladeshi students: Subject selection reasons and learning environment as potential risk factors," *Perspect. Psychiatr. Care*, vol. 57, no. 3, pp. 1150–1162, Jul. 2021, doi: 10.1111/ppc.12670.
- [5] M. A. Mamun *et al.*, "Prevalence and Psychiatric Risk Factors of Excessive Internet Use among Northern Bangladeshi Job-Seeking Graduate Students: A Pilot Study," *Int. J. Ment. Health Addict.*, vol. 19, no. 4, pp. 908–918, Aug. 2021, doi: 10.1007/s11469-019-00066-5.
- [6] S. C. Campisi, A. Ataullahjan, J.-A. B. Baxter, P. Szatmari, and Z. A. Bhutta, "Mental health interventions in adolescence," *Curr. Opin. Psychol.*, vol. 48, p. 101492, Dec. 2022, doi: 10.1016/j.copsyc.2022.101492.
- [7] Md. M. A. Shah, S. Ahmed, and S. M. Y. Arafat, "Demography and Risk Factors of Suicide in Bangladesh: A Six-Month Paper Content Analysis," *Psychiatry J.*, vol. 2017, pp. 1–5, 2017, doi: 10.1155/2017/3047025.
- [8] V. Ajdacic-Gross, "The Prevention of Mental Disorders has a Bright Future," *Front. Public Health*, vol. 2, Jun. 2014, doi: 10.3389/fpubh.2014.00060.
- [9] G. Serafini *et al.*, "Differential clinical characteristics and possible predictors of bipolarity in a sample of unipolar and bipolar inpatients," *Psychiatry Res.*, vol. 270, pp. 1099–1104, Dec. 2018, doi: 10.1016/j.psychres.2018.06.041.
- [10] D. Burns, H. Westra, M. Trockel, and A. Fisher, "Motivation and Changes in Depression," *Cogn. Ther. Res.*, vol. 37, no. 2, pp. 368–379, Apr. 2013, doi: 10.1007/s10608-012-9458-3.
- [11] J. Upton, "Beck Depression Inventory (BDI)," in *Encyclopedia of Behavioral Medicine*, M. D. Gellman, Ed., Cham: Springer International Publishing, 2020, pp. 202–203. doi: 10.1007/978-3-030-39903-0_441.
- [12] R. Sharp, "The Hamilton Rating Scale for Depression," *Occup. Med.*, vol. 65, no. 4, pp. 340–340, Jun. 2015, doi: 10.1093/occmed/kqv043.
- [13] W. W. K. Zung, "Zung Self-Rating Depression Scale and Depression Status Inventory," in *Assessment of Depression*, N. Sartorius and T. A. Ban, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 1986, pp. 221–231. doi: 10.1007/978-3-642-70486-4_21.
- [14] J. A. Yesavage *et al.*, "Development and validation of a geriatric depression screening scale: A preliminary report," *J. Psychiatr. Res.*, vol. 17, no. 1, pp. 37–49, Jan. 1982, doi: 10.1016/0022-3956(82)90033-4.
- [15] S. J. Brooks, S. P. Krulwicz, and S. Kutcher, "The Kutcher Adolescent Depression Scale: Assessment of Its Evaluative Properties over the Course of an 8-Week Pediatric Pharmacotherapy Trial," *J. Child Adolesc. Psychopharmacol.*, vol. 13, no. 3, pp. 337–349, Sep. 2003, doi: 10.1089/104454603322572679.
- [16] S. A. Montgomery and M. Åsberg, "A New Depression Scale Designed to be Sensitive to Change," *Br. J. Psychiatry*, vol. 134, no. 4, pp. 382–389, Apr. 1979, doi: 10.1192/bjp.134.4.382.
- [17] J. L. Cox, J. M. Holden, and R. Sagovsky, "Detection of Postnatal Depression: Development of the 10-item Edinburgh Postnatal Depression Scale," *Br. J. Psychiatry*, vol. 150, no. 6, pp. 782–786, Jun. 1987, doi: 10.1192/bjp.150.6.782.
- [18] P. Bech, N.-A. Rasmussen, L. R. Olsen, V. Noerholm, and W. Abildgaard, "The sensitivity and specificity of the Major Depression Inventory, using the Present State Examination as the index of diagnostic validity," *J. Affect. Disord.*, vol. 66, no. 2–3, pp. 159–164, Oct. 2001, doi: 10.1016/S0165-0327(00)00309-8.
- [19] S. Gilbody, T. Sheldon, and A. House, "Screening and case-finding instruments for depression: a meta-analysis," *Can. Med. Assoc. J.*, vol. 178, no. 8, pp. 997–1003, Apr. 2008, doi: 10.1503/cmaj.070281.
- [20] L. Costantini *et al.*, "A breakthrough in research on depression screening: from validation to efficacy studies," *Acta Biomed. Atenei Parm.*, vol. 92, no. 3, p. e2021215, Jul. 2021, doi: 10.23750/abm.v92i3.11574.
- [21] S. D'Alfonso, "AI in mental health," *Curr. Opin. Psychol.*, vol. 36, pp. 112–117, Dec. 2020, doi: 10.1016/j.copsyc.2020.04.005.
- [22] S. Aleem, N. U. Huda, R. Amin, S. Khalid, S. S. Alshamrani, and A. Alshehri, "Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions," *Electronics*, vol. 11, no. 7, p. 1111, Mar. 2022, doi: 10.3390/electronics11071111.
- [23] C. Chu, A.-L. Hsu, K.-H. Chou, P. Bandettini, and C. Lin, "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images," *NeuroImage*, vol. 60, no. 1, pp. 59–70, Mar. 2012, doi: 10.1016/j.neuroimage.2011.11.066.
- [24] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.
- [25] V. Conceição, I. Rothes, and R. Gusmão, "The association between stigmatizing attitudes towards depression and help seeking attitudes in college students," *PLOS ONE*, vol. 17, no. 2, p. e0263622, Feb. 2022, doi: 10.1371/journal.pone.0263622.
- [26] Md. S. Zulfiker, N. Kabir, A. A. Biswas, T. Nazneen, and M. S. Uddin, "An in-depth analysis of machine learning approaches to predict depression," *Curr. Res. Behav. Sci.*, vol. 2, p. 100044, Nov. 2021, doi: 10.1016/j.crbeha.2021.100044.

- [27] J. Cvetković, "Breast Cancer Patients' Depression Prediction by Machine Learning Approach," *Cancer Invest.*, vol. 35, no. 8, pp. 569–572, Sep. 2017, doi: 10.1080/07357907.2017.1363892.
- [28] C. Lee and H. Kim, "Machine learning-based predictive modeling of depression in hypertensive populations," *PLOS ONE*, vol. 17, no. 7, p. e0272330, Jul. 2022, doi: 10.1371/journal.pone.0272330.
- [29] A. A. Choudhury, Md. R. H. Khan, N. Z. Nahim, S. R. Tulon, S. Islam, and A. Chakrabarty, "Predicting Depression in Bangladeshi Undergraduates using Machine Learning," in *2019 IEEE Region 10 Symposium (TENSYP)*, Kolkata, India: IEEE, Jun. 2019, pp. 789–794. doi: 10.1109/TENSYP46218.2019.8971369.
- [30] R. Siddiqua, N. Islam, J. F. Bolaka, R. Khan, and S. Momen, "AIDA: Artificial intelligence based depression assessment applied to Bangladeshi students," *Array*, vol. 18, p. 100291, Jul. 2023, doi: 10.1016/j.array.2023.100291.
- [31] C. M. Hutton, L. W. Paton, D. McMillan, J. Cussens, S. Gilbody, and P. A. Tiffin, "Predicting persistent depressive symptoms in older adults: A machine learning approach to personalised mental healthcare," *J. Affect. Disord.*, vol. 246, pp. 857–860, Mar. 2019, doi: 10.1016/j.jad.2018.12.095.
- [32] S. Jiménez-Serrano, S. Tortajada, and J. M. García-Gómez, "A Mobile Health Application to Predict Postpartum Depression Based on Machine Learning," *Telemed. E-Health*, vol. 21, no. 7, pp. 567–574, Jul. 2015, doi: 10.1089/tmj.2014.0113.
- [33] A. Sau and I. Bhakta, "Screening of anxiety and depression among seafarers using machine learning technology," *Inform. Med. Unlocked*, vol. 16, p. 100228, 2019, doi: 10.1016/j.imu.2019.100228.
- [34] T. R. Shaha and M. Begum, "Prediction of Depression Severity via Feature Grouping and Machine Learning with Burn Depression Checklist," in *Applied Intelligence for Healthcare Informatics*, 1st ed., Boca Raton: CRC Press, 2025, pp. 85–99. doi: 10.1201/9781003583363-7.
- [35] K. A. Adamson and S. Prion, "Reliability: Measuring Internal Consistency Using Cronbach's α ," *Clin. Simul. Nurs.*, vol. 9, no. 5, pp. e179–e180, May 2013, doi: 10.1016/j.ecns.2012.12.001.
- [36] P. Sedgwick, "Spearman's rank correlation coefficient," *BMJ*, p. g7327, Nov. 2014, doi: 10.1136/bmj.g7327.
- [37] S. Arndt, C. Turvey, and N. C. Andreasen, "Correlating and predicting psychiatric symptom ratings: Spearman's r versus Kendall's tau correlation," *J. Psychiatr. Res.*, vol. 33, no. 2, pp. 97–104, Mar. 1999, doi: 10.1016/S0022-3956(98)90046-2.
- [38] A. M. Priyatno and T. Widiyaningtyas, "A SYSTEMATIC LITERATURE REVIEW: RECURSIVE FEATURE ELIMINATION ALGORITHMS," *JITK J. Ilmu Pengetah. Dan Teknol. Komput.*, vol. 9, no. 2, pp. 196–207, Feb. 2024, doi: 10.33480/jitk.v9i2.5015.
- [39] M. D. Tislenko, A. V. Gaidel, and A. V. Kupriyanov, "Comparison of feature selection algorithms for Data classification problems," in *2022 VIII International Conference on Information Technology and Nanotechnology (ITNT)*, Samara, Russian Federation: IEEE, May 2022, pp. 1–5. doi: 10.1109/ITNT55410.2022.9848765.
- [40] S. Shakeela, N. S. Shankar, P. M. Reddy, T. K. Tulasi, and M. M. Koneru, "Optimal Ensemble Learning Based on Distinctive Feature Selection by Univariate ANOVA-F Statistics for IDS," *Int. J. Electron. Telecommun.*, pp. 267–275, May 2021, doi: 10.24425/ijet.2021.135975.
- [41] H. Zhou, X. Wang, and R. Zhu, "Feature selection based on mutual information with correlation coefficient," *Appl. Intell.*, vol. 52, no. 5, pp. 5457–5474, Mar. 2022, doi: 10.1007/s10489-021-02524-x.
- [42] D. Colledani, P. Anselmi, and E. Robusto, "Machine learning-decision tree classifiers in psychiatric assessment: An application to the diagnosis of major depressive disorder," *Psychiatry Res.*, vol. 322, p. 115127, Apr. 2023, doi: 10.1016/j.psychres.2023.115127.
- [43] M. R. Gohari *et al.*, "Using random forest to identify correlates of depression symptoms among adolescents," *Soc. Psychiatry Psychiatr. Epidemiol.*, vol. 59, no. 11, pp. 2063–2071, Nov. 2024, doi: 10.1007/s00127-024-02695-1.
- [44] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, Madurai, India: IEEE, May 2019, pp. 1255–1260. doi: 10.1109/ICCS45141.2019.9065747.
- [45] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, Jan. 2002, doi: 10.1023/A:1012487302797.
- [46] R. P. Sonjaya, A. R. Gintara, L. S. Riza, M. Nursalman, E. Nugraha, and D. Wahyudin, "Predicting Student Depression Using the Naive Bayes Model on the Student Depression Dataset from Kaggle," *JENTIK J. Pendidik. Teknol. Inf. Dan Komun.*, vol. 4, no. 1, pp. 9–20, Jun. 2025, doi: 10.58723/jentik.v4i1.448.
- [47] M. Skurichina and R. P. W. Duin, "Bagging, Boosting and the Random Subspace Method for Linear Classifiers," *Pattern Anal. Appl.*, vol. 5, no. 2, pp. 121–135, Jun. 2002, doi: 10.1007/s100440200011.
- [48] A. I. Naimi and L. B. Balzer, "Stacked generalization: an introduction to super learning," *Eur. J. Epidemiol.*, vol. 33, no. 5, pp. 459–464, May 2018, doi: 10.1007/s10654-018-0390-z.
- [49] I. Ahmed, A. Brahmacharimayum, R. H. Ali, T. A. Khan, and M. O. Ahmad, "Explainable AI for Depression Detection and Severity Classification From Activity Data: Development and Evaluation Study of an Interpretable Framework," *JMIR Ment. Health*, vol. 12, pp. e72038–e72038, Sep. 2025, doi: 10.2196/72038.
- [50] A. A. Jo, E. Deni Raj, A. S. Vino, and P. V. Menon, "Exploring Explainable AI for Enhanced Depression Prediction in Mental Health," in *2024 First International Conference on Innovations in Communications, Electrical and Computer Engineering (ICICEC)*, Davangere, India: IEEE, Oct. 2024, pp. 1–7. doi: 10.1109/ICICEC62498.2024.10808414.

BIOGRAPHY OF AUTHORS

Tumpa Rani Shaha is a researcher in Computer Science and Engineering at Dhaka University of Engineering Technology (DUET), Bangladesh. She received her B.Sc and M.Sc degrees under department of Computer Science and Engineering from same university. Currently, she is a faculty member at the Department of Computer Science and Engineering, Gopalganj Science and Technology University, Gopalganj, Bangladesh. Her research areas include Causal Machine Learning, Trustworthy AI, Human Computer Interaction, and Mental Health. email: tumpa,cse49@gmail.com.



MOMOTAZ BEGUM PhD, is a distinguished professor in the Department of Computer Science and Engineering at DUET, Gazipur. Her research interests encompass Software Reliability, Software Engineering, Artificial Neural Network, Data Mining, Data Clustering, Information System, and Analysis. Dr. Begum's specialization lies in Software Rejuvenation and Software Aging, as well as Human Pose Recognition. She received a Ph.D. degree (Information Engineering) from Hiroshima University, Japan, fully funded by the Japanese government, specifically the Ministry of Education, Culture, Sports, Science and Technology (Monbukagakusho (MEXT)) in Oct 2014-Sep 2017, and an MSc in CSE from, DUET, Gazipur, Bangladesh, in 2013. She has published several journal and conference papers in the ,world's reputed journals and conferences on the different topics of ctoputer science, such as IoT, machine learning, neural networks. Email: drmomotaz@duet.ac.bd