Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00098-8

Original Article

Model-Agnostic Interpretability for Student Depression Screening: Integrating XAI and Multi-Criteria Model Selection

Mohima Binte Rasel [1] , Md. Abid Hasan Rafi [1, ✉] , Mst. Fatematuj Johora [1] , Pankaj Bhowmik [2]

1. *Electronics and Communication Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur, BGD*
2. *Computer Science and Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur, BGD*

Received: April 21, 2026 | Review began: May 13, 2026 | Review ended: May 25, 2026 | Published: June 03, 2026

Abstract

Depression in university students has emerged as a critical public health issue, necessitating scalable and comprehensible methodologies for early identification. This research introduces an Explainable Artificial Intelligence framework to forecast depression severity among Bangladeshi undergraduate students aged 18 to 28 years, utilizing sociodemographic, lifestyle, behavioral, and psychometric characteristics obtained from validated instruments such as the Beck Depression Inventory, Second edition, Patient Health Questionnaire 9, Center for Epidemiologic Studies-Depression scale, and the University of California Los Angeles Loneliness Scale. We train tree-based machine learning models like Decision Tree, Random Forest, XGBoost, and LightGBM on a 27-dimensional feature space and then test them using a wide range of metrics that measure predictive performance, calibration, stability, fairness, and interpretability. The results show that Decision Tree and LightGBM have the highest classification accuracy (96.0%) and the best ability to tell the difference between classes. XGBoost, on the other hand, gives the best calibrated probability estimates. SHapley Additive exPlanation-based explainability analysis shows that composite psychometric scores are the most important predictors of severe depression. This is backed up by large effect sizes and statistically significant differences between groups. The explanations are very stable across subsamples, and the fairness evaluation shows that there is very little bias between gender groups. These results show that combining accurate machine learning models with strong interpretability and fairness assessment can be a dependable and clear way to screen for depression in academic settings with limited resources.

Categories: Explainable AI, Health Informatics, Machine Learning (ML)

Keywords: explainable ai, depression prediction, shap, mental health, lime

Introduction

Mental health disorders are a significant and inadequately addressed concern in higher education, impacting over 264 million individuals worldwide, with university students exhibiting greater susceptibility than the general populace. Academic stressors such as heavy workloads, financial difficulties, social isolation, poor sleep, and increasing independence contribute significantly to this risk [1,2,3]. Globally, prevalence remains high, with substantial rates of depression, anxiety, stress, sleep disorders, and suicidal ideation among students [4]. The COVID-19 pandemic further intensified these challenges, significantly worsening psychological distress [5] across university populations worldwide [6,7]. In Bangladesh, a large proportion of students experience anxiety, depression, or panic symptoms, primarily linked to academic pressure [8]. Broader evidence also highlights considerable prevalence of anxiety and depression, with contributing lifestyle factors such as excessive screen time and poor diet [9]. A dataset has also been developed to better understand student mental health conditions in the Bangladeshi context [10].

Despite widespread mental health issues, only a minority of students seek professional treatment [11,12]. In particular, first-generation students demonstrate lower help-seeking behavior compared to continuing-generation peers [13]. Multiple structural barriers exist, including stigma, cost, limited counseling access, and weak institutional response systems [14]. Studies further show that mental health challenges [15] among university students are influenced by sleep quality, physical inactivity, and academic pressure, with higher vulnerability observed in marginalized groups [16,17]. Additional evidence highlights the impact of adverse childhood experiences and increased risk among Lesbian, Gay, Bisexual Transgender, or Queer plus other sexual orientations and gender identity (LGBTQ+) students [18,19].

Student mental health is also strongly influenced by sleep patterns and behavioral health factors, with poor sleep quality significantly associated with psychological distress [20]. Behavioral data-driven studies further suggest that daily digital activity patterns can be predictive of mental health status [21]. A comprehensive synthesis of global evidence confirms the high burden of mental disorders among university students [22]. Qualitative research also highlights lived experiences of students facing mental health challenges on campus [23].

Economic and academic stressors further contribute to depression and anxiety risk, with significant associations reported in multiple studies [24]. Mental health services across universities vary widely, and institutional support remains insufficient in many settings [25]. Physical activity has been shown to play a protective role in student mental health outcomes [26]. The COVID-19 pandemic also significantly worsened mental health across student populations, particularly among vulnerable demographic groups [27]. Recent cross-sectional evidence continues to confirm high levels of psychological distress during and after the pandemic period [28]. Comparisons between undergraduate and graduate populations further show differences in mental health burden [29]. Research in specialized populations such as nursing students has also increased, focusing on stress, resilience, and psychological well-being trends [30].

Machine learning approaches have recently been explored for predicting mental health outcomes using behavioral and psychometric data; however, many existing studies primarily emphasize predictive accuracy while providing limited attention to interpretability, fairness, calibration, and model stability [31,32]. Furthermore, explainable artificial intelligence (XAI)-integrated mental health prediction frameworks remain underexplored, particularly among university students in developing countries such as Bangladesh. These limitations reduce the transparency, reliability, and practical applicability of machine learning systems in real-world mental health screening environments.

To address these gaps, this study develops and evaluates a model-agnostic XAI framework for depression severity prediction among Bangladeshi university students using tree-based machine learning models [33,34], including Decision Tree, Random Forest, XGBoost, and LightGBM. The study investigates which model provides the best trade-off between predictive accuracy, calibration, fairness, stability, and interpretability using 27 behavioral and psychometric features. SHapley Additive exPlainations (SHAP)-based explainability [35] is employed for both global and local interpretation of model decisions. Experimental results demonstrate strong predictive capability, with Decision Tree and LightGBM achieving the highest classification accuracy, while XGBoost provides superior probabilistic calibration performance. Additionally, psychometric indicators such as Beck Depression Inventory (BDI) and Center for Epidemiologic Studies Depression Scale (CES-D) scores emerge as dominant predictors of depression severity. The study introduces a pipeline designed to support scalable and transparent mental health screening in resource-constrained university environments.

Materials And Methods

Dataset and survey instruments

The dataset utilized in this study was obtained from the publicly available Mendeley Data repository [10]. It consists of organized survey responses collected from 500 undergraduate students. The participants' ages ranged from 18 to 28 years, with a mean age of 22.28 years and a standard deviation of 2.1. Data collection was conducted over a 3-month period using both online and paper-based survey instruments. The study protocol received institutional review board approval, and written informed consent was obtained from all participants prior to data collection.

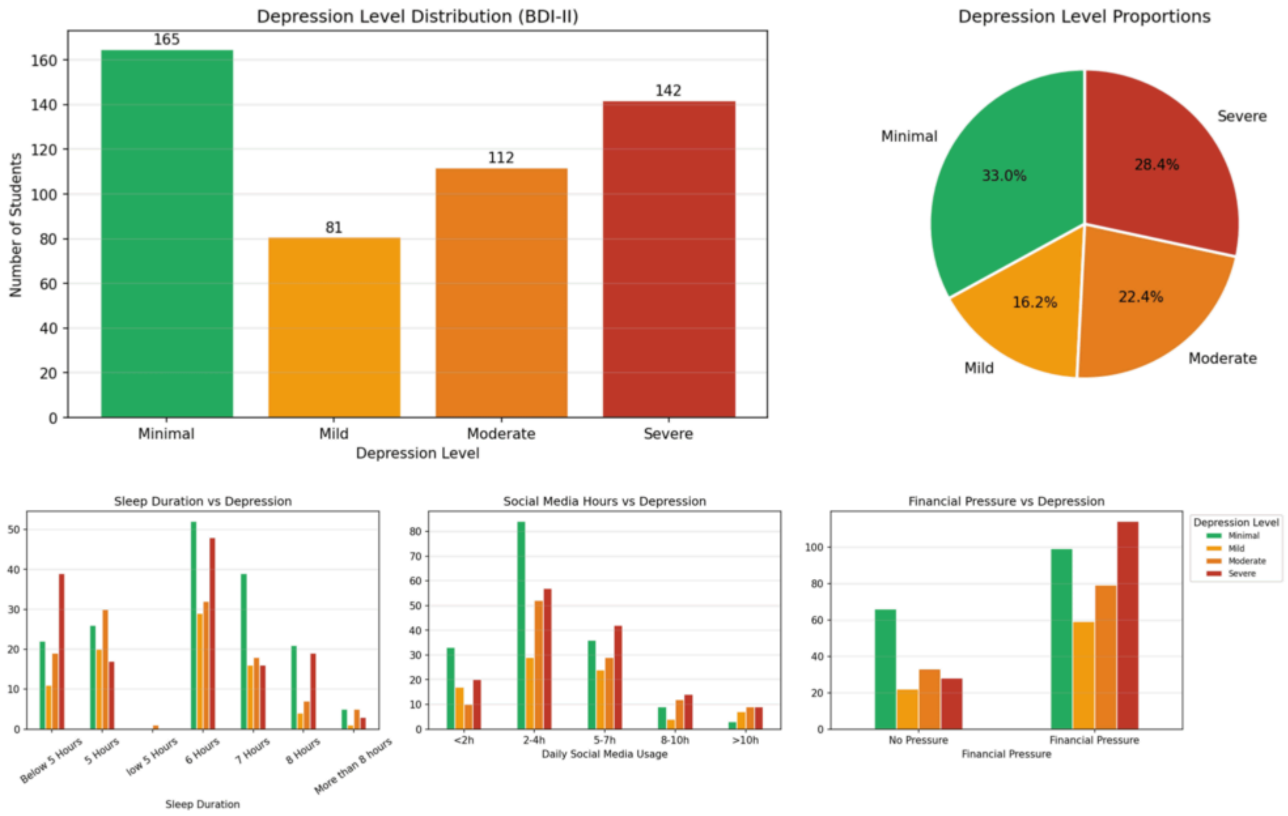

FIGURE 1: Exploratory data analysis of the undergraduate mental health dataset (n = 500), showing depression class distribution (counts and proportions) and class-wise relationships with sleep duration, social media usage, and financial pressure.

The visualization highlights relative class balance and clinically meaningful lifestyle and socioeconomic gradients across severity levels. BDI-II, Beck Depression Inventory, Second Edition.

Figure *1* shows that the four target classes are distributed in a way that is suitable for multiclass modeling without extreme imbalance penalties. The survey comprised 63 items organized across four validated psychometric instruments and a set of sociodemographic and lifestyle questions. The BDI, Second Edition (BDI-II), consists of 21 items (20 used in this study after item consolidation) that assess affective, cognitive, and somatic symptoms of depression over the past 2 weeks. Each item carries a 0-3 ordinal score, producing a composite BDI Total score in the range, which is subsequently mapped to four categorical depression levels: Minimal (0-13), Mild (14-19), Moderate (20-28), and Severe (29-60). The Patient Health Questionnaire (PHQ-9) contributes three supplementary items. The Center for Epidemiologic Studies Depression Scale (CES-D) consists of 9 items scored on a 4-point frequency scale. The University of California, Los Angeles (UCLA) Loneliness Scale (Version 3, abbreviated to 8 items) assesses subjective loneliness on a four-point frequency scale. The exploratory analysis supports an initial quality check of both class structure and covariate behavior before supervised learning.

Feature construction and engineering

After preprocessing, the 63 raw survey items were transformed into a structured 27-dimensional feature set for model development. Binary lifestyle-related variables, including physical activity, smoking, alcohol consumption, debt status, financial pressure, medication use, and workload condition, were converted into numerical binary representations to ensure compatibility with machine learning algorithms. Academic standing was encoded as an ordinal feature representing the progression from first to fourth year, while socioeconomic status was encoded across multiple levels ranging from lower to upper class. Nominal attributes such as relationship status and residential area were transformed using one-hot encoding to preserve category independence without introducing ordinal relationships. In addition to the original survey attributes, two derived categorical variables were introduced to better capture behavioral patterns associated with depression. Sleep duration was grouped into insufficient, normal, and excessive sleep categories, while daily social media usage was categorized into five intensity levels ranging from low to extremely high usage. A workload intensity feature was further designed by combining reported workload condition with academic standing, allowing the model to capture both the presence of workload pressure and its increasing impact across academic years. Four psychometric assessment scores, including BDI, UCLA Loneliness Scale, CES-D, and PHQ-based measures, were calculated by aggregating their respective questionnaire item responses. Prior to model training, all numerical features were standardized to ensure that each feature contributed proportionally during learning and to prevent scale-related bias. Missing values, which accounted for less than 3% of the non-target data, were handled using feature-wise median imputation to preserve dataset consistency and minimize information loss. Depression prediction was formulated as a supervised multi-class classification problem in which each student was assigned to one of the four depression severity categories: Minimal, Mild, Moderate, or Severe. The machine learning models were trained to learn meaningful relationships between the standardized feature representations and the corresponding depression classes while maintaining strong generalization performance on unseen data. In addition to class predictions, probabilistic outputs were generated by the classifiers, enabling further analyses related to model calibration, fairness assessment, and interpretability.

Proposed methodological workflow

The proposed workflow of this study is designed to avoid the common mistake of choosing models solely on their accuracy. As illustrated in Figure *2*, generated using Gemini 3.1 Pro [36], the model development pipeline begins with rigorous preprocessing controls followed by stratified validation to ensure robust and unbiased evaluation. Then, it moves on to a multi-axis assessment of fairness, discrimination, calibration, and interpretability. This staged design ensures that a model with good ranking ability, but poor confidence reliability or subgroup behavior is not promoted without further testing. The explainability layer is also built into the pipeline rather than added at the end. It allows to understand the quality of the predictions in relation to the quality of the explanations. Such integration is necessary in mental health screening settings, where model outputs must be both efficacious and clinically justifiable.

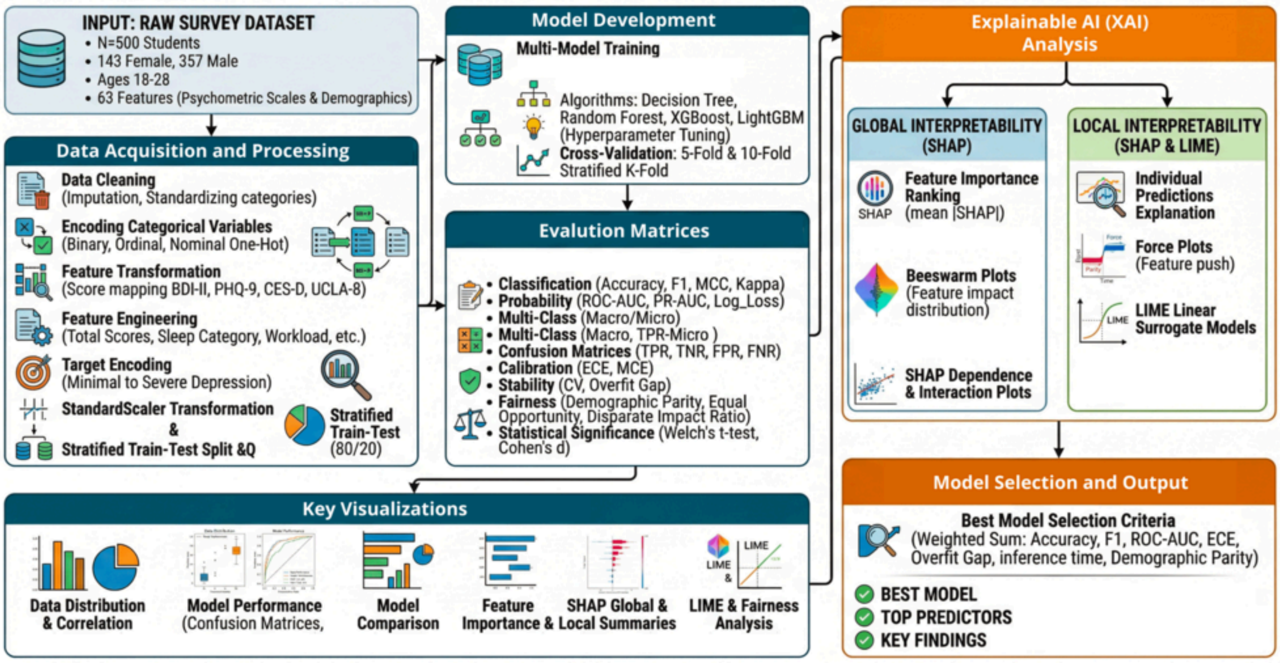

FIGURE 2: Proposed workflow from raw survey data to validated and interpretable model outputs, including preprocessing, feature engineering, stratified splitting, multi-model training, cross-validation, calibration and fairness assessment, and global/local explainability via SHAP and LIME.

SHAP, Shapley Additive exPlanations; LIME, Local Interpretable Model-Agnostic Explanations

Machine learning classifiers

All machine learning models were trained and evaluated using the same preprocessing pipeline, dataset splitting strategy, and scoring criteria to ensure a fair and unbiased comparison across models. This standardized experimental setup minimized model-specific bias and allowed performance differences to be attributed to the learning mechanisms of the algorithms themselves rather than inconsistencies in the evaluation procedure. The selected models represent different trade-offs relevant to real-world deployment, including interpretability, robustness, computational efficiency, and predictive power. Decision Tree was included because of its transparent and human-readable decision structure, making it easy to understand how predictions are generated. The model works by recursively dividing the feature space into smaller groups based on the features that best separate the depression classes, thereby creating a hierarchy of decision rules. Although Decision Trees are highly interpretable, they are often sensitive to noise and may suffer from overfitting when used individually.

To improve robustness and generalization, Random Forest was employed as an ensemble extension of Decision Trees in which multiple trees were trained using randomly sampled subsets of the training data and feature space, with final predictions determined through majority voting across all trees. This ensemble strategy reduced variance, improved prediction stability, and enhanced generalization compared to a single Decision Tree, while out-of-bag samples provided an internal mechanism for evaluating model performance on unseen data. XGBoost was selected for its strong predictive capability and effectiveness in structured tabular datasets, utilizing a boosting strategy where trees are trained sequentially to progressively correct the prediction errors of preceding trees. The model incorporates regularization techniques to control tree complexity and reduce overfitting while leveraging gradient-based optimization to efficiently identify informative feature splits and improve classification performance across the four depression severity classes. LightGBM was additionally employed as a computationally efficient gradient boosting framework optimized for faster training and lower memory consumption. Unlike conventional depth-wise tree growth methods, LightGBM expands tree leaf-wise by prioritizing branches that produce the greatest reduction in prediction error, often resulting in improved predictive accuracy. Furthermore, LightGBM integrates optimization strategies such as Gradient-based One-Side Sampling (GOSS), which emphasizes highly informative samples with larger prediction errors, and Exclusive Feature Bundling (EFB), which reduces feature dimensionality by combining sparse and mutually exclusive features, thereby enabling scalable and efficient learning for complex classification tasks.

Explainability with SHAP and LIME

Interpretability was treated as a fundamental analytical objective of this study rather than an additional post-processing component. To ensure both global and local understanding of the model's behavior, two complementary XAI techniques, SHAP and Local Interpretable Model-Agnostic Explanations (LIME), were employed. SHAP was used to identify the overall factors contributing to depression risk across the entire dataset by quantifying the contribution of each feature to the model's predictions. Grounded in cooperative game theory, SHAP provides theoretically consistent and reliable feature attributions, enabling the identification and ranking of the most influential factors associated with depression severity. In this study, SHAP analysis was conducted across all four depression classes, with particular emphasis on the Severe Depression category, while the TreeSHAP implementation enabled efficient computation of feature contributions for tree-based ensemble models. In contrast, LIME was employed to explain individual predictions by approximating the complex model locally around a specific student instance using a simpler and interpretable surrogate model. Through local perturbation analysis, LIME identifies the most influential features responsible for a particular prediction while maintaining explanation sparsity for improved human readability. Together, SHAP and LIME provided a comprehensive interpretability framework in which SHAP explained what generally drives depression risk across the population, whereas LIME clarified why a specific student received a particular predicted class, thereby improving the transparency, reliability, and trustworthiness of the proposed machine learning framework.

Methodological considerations and limitations

The proposed research process is systematic, reproducible, and follows a well-structured experimental pipeline, but there are a few points to consider. While stratified validation and consistent preprocessing across models were performed, the study is based on a single publicly available dataset of university students in Bangladesh, potentially limiting the generalizability of the results to other cultural or institutional settings. The survey data are cross-sectional, which limits causal interpretation of the relationships between features and depression severity, allowing only associative interpretation. Also, various evaluation metrics, such as stability, fairness, explainability, and accuracy, were integrated to provide a comprehensive view. Still, this multi-objective evaluation may involve compromises that cannot be fully optimized in a single, unified framework. Moreover, while SHAP and LIME are both very interpretable, they are both post-hoc explanation methods and may not accurately reflect all aspects of the model decision-making process. Despite these limitations, the experimental design remains robust, and the use of multiple complementary models strengthens the reliability of the comparative analysis. The study's findings generally support the conclusions, as all are directly supported by empirical evaluation using various metrics and validation techniques.

Results And Discussion

This section presents the experimental results in a structured and reproducible manner, beginning with discrimination performance and subsequently examining calibration, fairness, and interpretability to facilitate the selection of a model that balances predictive accuracy, probabilistic reliability, equity, and explainability. To ensure transparency and reproducibility of the research, the complete implementation notebook, including data preprocessing, model training, evaluation procedures, and explainability analyses, has been publicly shared [37]. The provided notebook enables independent verification of the experimental workflow and supports the reproducibility of the reported findings.

How to cite this article:
Rasel M, Rafi M, Johora M, et al. (June 03, 2026) Model-Agnostic Interpretability for Student Depression Screening: Integrating XAI and Multi-Criteria Model Selection. Cureus J Comput Sci 3 : es44389-026-00098-8. DOI https://doi.org/10.7759/s44389-026-00098-8

Experimental setup

The proposed experimental framework was implemented in a Python-based environment using Google Colaboratory to ensure reproducibility and ease of execution. The implementation relied on widely used scientific computing and machine learning libraries, including NumPy, Pandas, Matplotlib, and Seaborn for data manipulation and visualization, while scikit-learn was used to develop the overall machine learning workflow. Advanced ensemble and boosting-based models were implemented using XGBoost and LightGBM, and model interpretability was supported through SHAP and LIME libraries. To ensure reproducibility, a fixed random seed (42) was applied throughout all experiments. The dataset was loaded directly into the Colab environment via the built-in file upload interface, consisting of 501 samples and 63 features, and initial inspection was performed using Pandas to verify data integrity prior to preprocessing and model training. For model development, multiple classification algorithms were evaluated, including Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM, along with stacking-based ensemble learning to enhance predictive performance. The training and evaluation pipeline employed stratified train-test splitting and cross-validation techniques, and model performance was assessed using a comprehensive set of metrics, including Accuracy, Precision, Recall, F1-score, Matthews Correlation Coefficient (MCC), Cohen's Kappa, Receiver Operating Characteristic Area Under the Curve (ROC-AUC), and log-loss. In addition, calibration analysis and confusion matrix visualization were used to evaluate model reliability and predictive behavior, while stratified k-fold cross-validation was applied to ensure robustness and generalization. All experiments were executed on the Google Colab runtime environment, leveraging cloud-based computational resources for efficient training and evaluation.

Classification performance

Table *1* reports classification results on the held-out 20% test set (n=100). Decision Tree and LightGBM achieved the highest accuracy (96.00%), followed by Random Forest and XGBoost (95.00%). Logistic Regression, despite inherent interpretability, reached 83.00%, indicating that nonlinear boundaries are important in four-class severity prediction. Weighted F1 tracked accuracy closely (Decision Tree: 0.9605; LightGBM: 0.9593). MCC, which is robust to class imbalance, confirmed the same ranking: Decision Tree (0.9460), LightGBM (0.9455), Random Forest (0.9336), XGBoost (0.9321), and Logistic Regression (0.7744).

How to cite this article:
Rasel M, Rafi M, Johora M, et al. (June 03, 2026) Model-Agnostic Interpretability for Student Depression Screening: Integrating XAI and Multi-Criteria Model Selection. Cureus J Comput Sci 3 : es44389-026-00098-8. DOI https://doi.org/10.7759/s44389-026-00098-8

Metric	LightGBM	Decision Tree	Random Forest	XGBoost
Accuracy	0.9600	0.9600	0.9500	0.9500
Balanced Accuracy	0.9460	0.9558	0.9359	0.9304
Precision (Weighted)	0.9595	0.9625	0.9564	0.9496
Recall (Weighted)	0.9600	0.9600	0.9500	0.9500
F1 (Weighted)	0.9593	0.9605	0.9498	0.9487
F2 (Weighted)	0.9596	0.9600	0.9491	0.9492
F0.5 (Weighted)	0.9593	0.9615	0.9530	0.9489
MCC	0.9455	0.9460	0.9336	0.9321
Cohen's kappa	0.9453	0.9455	0.9316	0.9314

TABLE 1: Classification metrics on the 20% held-out test set (n = 100).

MCC, Matthews Correlation Coefficient

When model outputs could help with clinical triage, class-wise error structure is more useful than overall accuracy. Figure *3* shows that most of the predictions are on the main diagonal, which means that the four severity classes are stable and separate. Residual errors are primarily found between adjacent levels, especially Mild and Moderate, which is anticipated when symptom boundaries are continuous instead of sharply defined. The boosted models have a stronger hold on severe cases, which is consistent with their higher probabilistic quality in later evaluations.

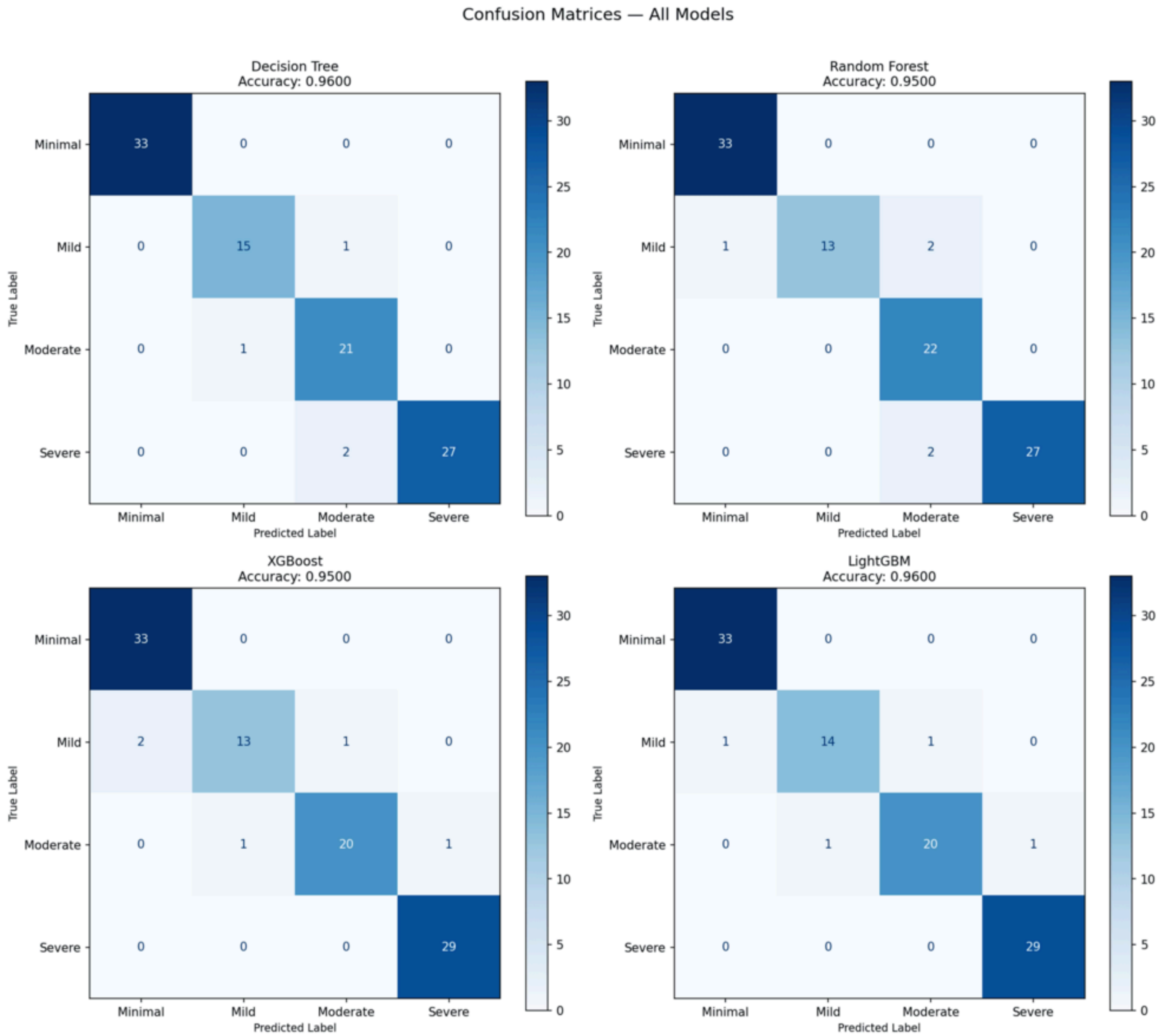

FIGURE 3: Confusion matrices for Decision Tree, Random Forest, XGBoost, and LightGBM on the held-out test set.

Diagonal cells indicate correct predictions, while off-diagonal cells reveal class-specific confusion, primarily between adjacent severity categories.

Probability calibration and ranking metrics

XGBoost achieves the strongest probabilistic performance, with the lowest Brier Score (0.0131) and Log Loss (0.0816), followed closely by LightGBM (Brier = 0.0149, Log Loss = 0.0957), as summarized in Table *2*, while both models maintain near-perfect ranking performance and strong calibration [Expected Calibration Error (ECE) = 0.0185 and 0.0205). In contrast, Decision Tree, despite high accuracy, shows weaker probability quality with lower Precision-Recall Area Under the Curve (PR-AUC) (0.9499) and much higher Log Loss (0.7882), whereas Random Forest exhibits notable miscalibration (ECE = 0.1061) and Logistic Regression demonstrate moderate calibration (0.0628) with comparatively lower discrimination.

Metric	LightGBM	Decision Tree	Random Forest	XGBoost
ROC-AUC (OvR, Weighted)	0.9986	0.9840	0.9952	0.9990
ROC-AUC (OvR, Macro)	0.9983	0.9843	0.9939	0.9986
PR-AUC (avg)	0.9931	0.9499	0.9768	0.9940
Log Loss	0.0957	0.7882	0.3663	0.0816
Brier Score (avg)	0.0149	0.0201	0.0420	0.0131
Gini Coefficient	0.9972	0.9680	0.9904	0.9980
ECE	0.0205	0.0235	0.1061	0.0185
MCE	0.6936	0.5173	0.4160	0.4652
Avg TPR	0.9460	0.9558	0.9359	0.9304
Avg TNR	0.9866	0.9874	0.9834	0.9828
Avg FPR	0.0134	0.0126	0.0166	0.0172
Avg FNR	0.0540	0.0442	0.0641	0.0696
CV5 Mean Accuracy	0.9480	0.9540	0.9520	0.9540
CV5 Std	0.0194	0.0196	0.0293	0.0265
Overfitting Gap	0.0400	0.0175	0.0450	0.0500

TABLE 2: Probabilistic, calibration, confusion-matrix, and stability metrics.

ROC-AUC, Receiver Operator Characteristic Area Under the Curve; PR-AUC, Precision-Recall Area Under the Curve; ECE, Expected Calibration Error; MCE, Maximum Calibrated Error; TPR, True Positive Rate; TNR, True Negative Rate; FPR, False Positive Rate; FNR, False Negative Rate

Threshold-dependent discrimination is consistently robust across all architectures; however, separation quality is not uniform across classes. Figure *4* shows that endpoint classes like Minimal and Severe are recovered with very steep curves and AUC behavior that is almost perfect. The ensemble boosting models produce smoother trajectories with fewer abrupt changes in threshold stability than the single-tree baseline, which is helpful when screening policies need to adjust the sensitivity-specificity trade-off. The smoother behavior indicates that ranking quality remains consistent across a broader range of operations, not just at a single cutoff. This kind of stability makes it more likely to be used in places where thresholds may differ across institutions or resource conditions.

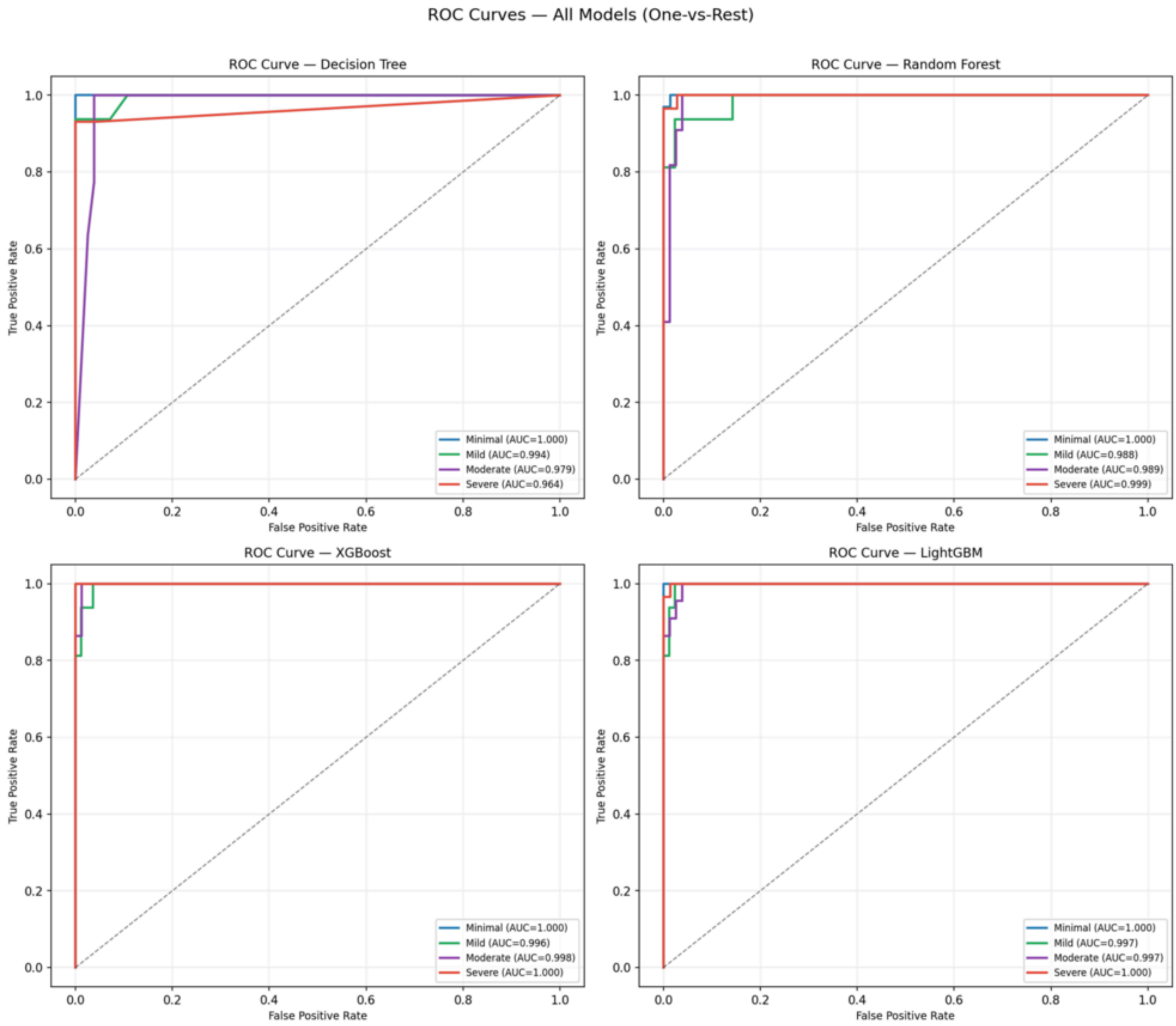

FIGURE 4: One-vs-rest ROC curves for all evaluated models and depression classes, illustrating class-discrimination behavior across thresholds.

Curves near the top-left corner indicate strong separability, with boosted ensembles showing near-perfect AUC profiles across most classes.

AUC, Area Under the Curve; ROC, Receiver Operating Characteristics

Threshold-dependent discrimination is consistently robust across all architectures; however, separation quality is not uniform across classes. Endpoint classes like Minimal and Severe are recovered with very steep curves and AUC behavior that is almost perfect. The ensemble boosting models produce smoother trajectories with fewer abrupt changes in threshold stability than the single-tree baseline, which is helpful when screening policies need to adjust the sensitivity-specificity trade-off. The smoother behavior indicates that ranking quality remains consistent across a broader range of operations, not just at a single cutoff. This kind of stability makes it more likely to be used in places where thresholds may differ across institutions or resource conditions.

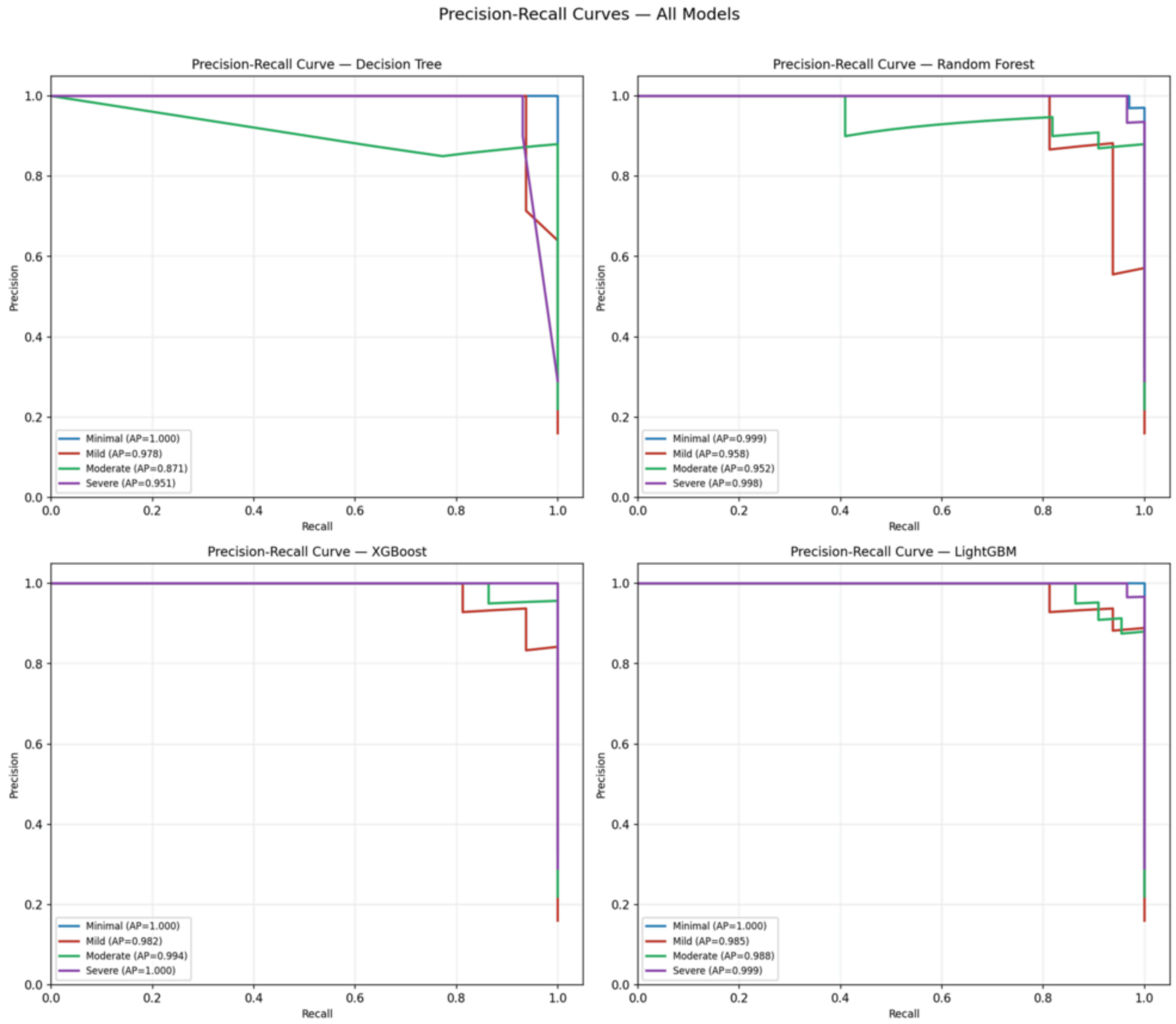

FIGURE 5: Precision-recall curves for one-vs-rest class evaluation.

These curves emphasize positive-class retrieval quality under class imbalance and show that boosted models maintain high precision over wide recall ranges for most depression categories.

Figure *5* shows that the best models keep high precision even when recall goes up in important classes. Endpoint categories are still the most stable, but the Mild and Moderate classes show earlier drops in precision, which makes sense given that the symptom profiles overlap only partially. The same mid-spectrum uncertainty was already clear in the confusion matrices, which showed that the complementary diagnostics were working together. This agreement across different points of view supports the idea that middle-severity ranges are harder to separate than boundary classes.

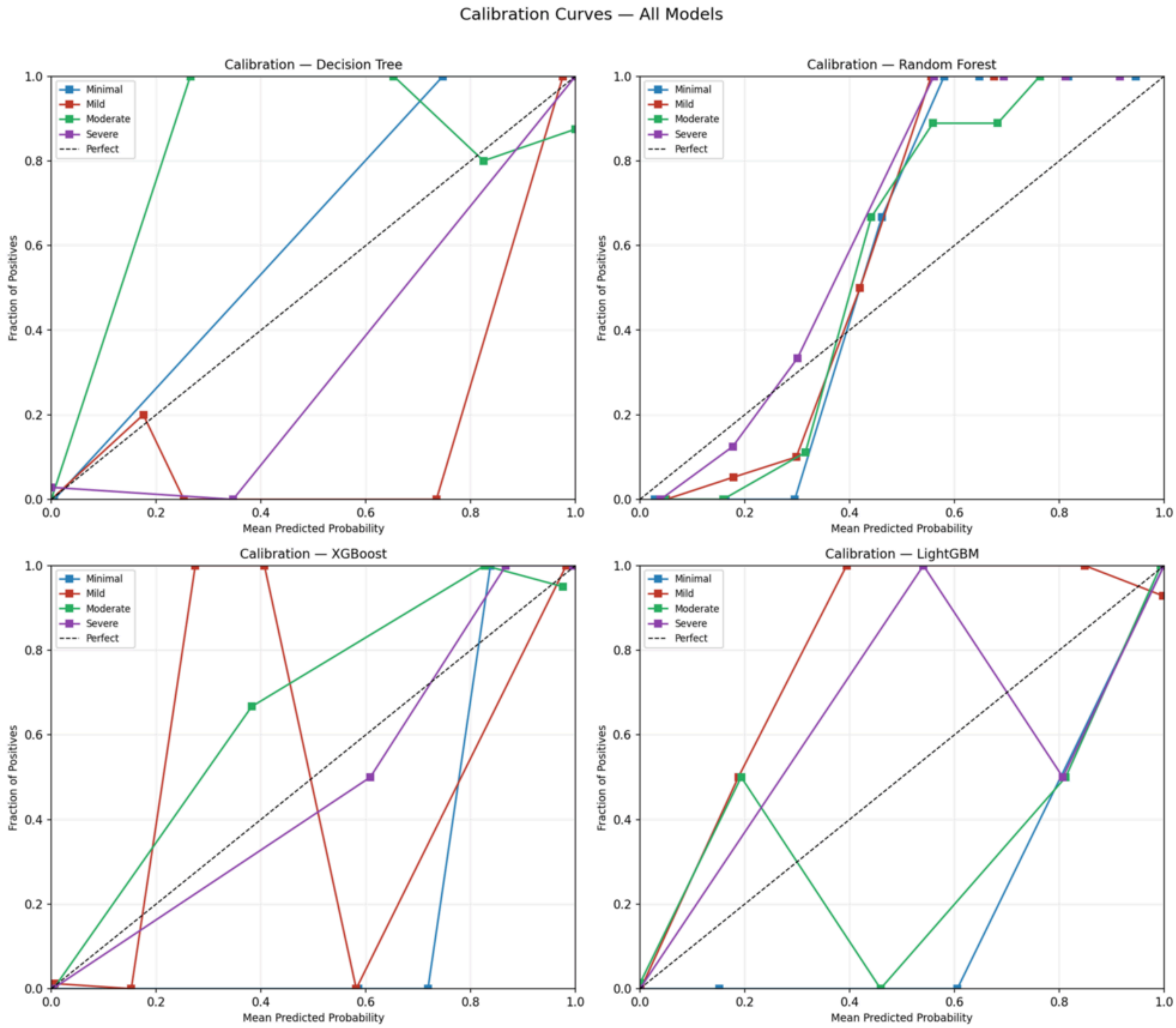

FIGURE 6: Reliability (calibration) diagrams comparing predicted probabilities with observed frequencies for each class and model.

The diagonal indicates perfect calibration; deviations indicate overconfidence or underconfidence in probabilistic outputs.

The calibration diagnostics in Figure 6 show that the quality of confidence changes depending on the architecture, even when discrimination stays high. Ensemble methods usually yield smoother, more realistic confidence profiles than single-tree estimates, which are more jagged in areas with low support. The boosted models still exhibit the best calibration across most classes, supporting their use in probability-based prioritization pipelines. These results support the importance of calibration-aware model selection for counseling triage situations.

Fairness analysis

Table 3 shows how fair the metrics are for men and women. The values for Demographic Parity Difference (DPD) were similar across models (0.219-0.246), which means that the predicted moderate-to-severe risk was similar across subgroups. The Equal Opportunity Difference (EOD), which shows the differences in true positive rate between subgroups, was much lower. Random Forest got an EOD of 0.000, which means that the sensitivity was the same for both male and female students. The Disparate Impact Ratio (DIR) was between 0.641 and 0.683, which is below the usual 0.8 level.

Model	DPD (↓)	EOD (↓)	DIR (↑)
LightGBM	0.2464	0.0294	0.6441
Decision Tree	0.2464	0.0294	0.6441
Random Forest	0.2193	0.0000	0.6832
XGBoost	0.2464	0.0294	0.6441

TABLE 3: Gender fairness metrics. DPD and EOD closer to 0 are more fair; DIR closer to 1 is more fair.

DPD, Demographic Parity Difference; EOD, Equal Opportunity Difference; DIR, Disparate Impact Ratio

The comparative profiles in Figure *7* demonstrate that prediction shapes are generally uniform across model families, suggesting that class-allocation behavior is influenced more by a common data signal than by instability unique to individual algorithms. Because the sample includes a larger group of men, differences in raw counts need to be examined alongside normalized parity metrics. Even with that unbalanced context, the models don't show sudden changes in how subgroups are assigned across architectures. This stability supports the fairness interpretation quantitatively detailed in Table *3*.

How to cite this article:
Rasel M, Rafi M, Johora M, et al. (June 03, 2026) Model-Agnostic Interpretability for Student Depression Screening: Integrating XAI and Multi-Criteria Model Selection. Cureus J Comput Sci 3 : es44389-026-00098-8. DOI https://doi.org/10.7759/s44389-026-00098-8

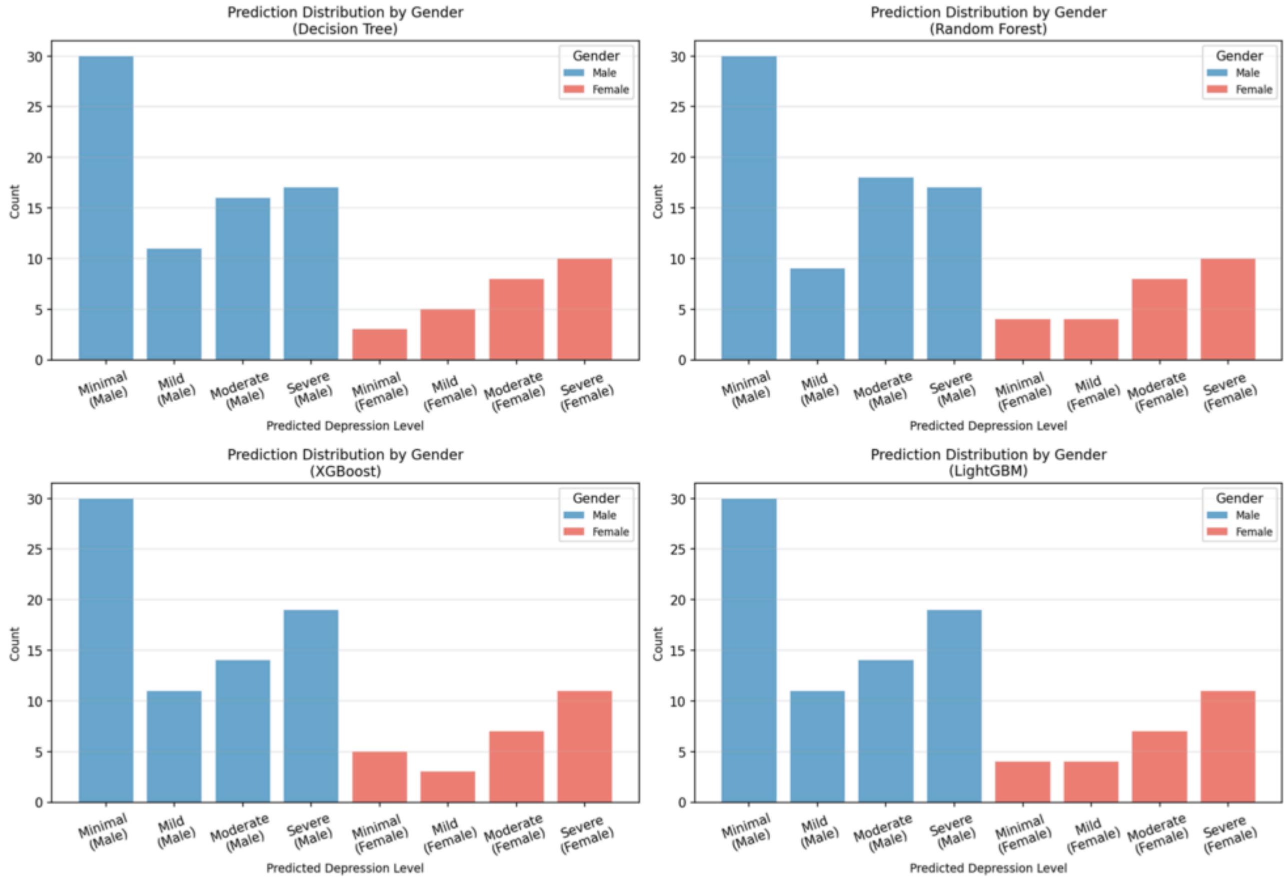

FIGURE 7: Model-wise distribution of predicted depression severity by gender.

The chart compares subgroup allocation patterns across Decision Tree, Random Forest, XGBoost, and LightGBM to support fairness diagnostics beyond aggregate parity metrics.

Statistical significance of key features

The information in Table 4 reports Welch's two-sample *t*-tests comparing Severe ($y = 3$) and Non-Severe ($y < 3$) groups across seven key features. BDI Total showed the strongest separation (Cohen's $d = 3.39$; 95% CI: [23.55, 26.35]; $p < 0.0001$), followed by CES-D Total ($d = 2.25$; 95% CI: [9.10,10.94]). UCLA Loneliness also showed a large effect ($d = 0.82$, $p < 0.0001$). Financial Pressure was significant ($p = 0.0009$) but with a small effect ($d = 0.32$). Sleep Category, Social Media Usage, and Age were not significant at $\alpha = 0.05$, suggesting that their predictive contribution is more likely interaction-driven than strongly bivariate.

How to cite this article:
Rasel M, Rafi M, Johora M, et al. (June 03, 2026) Model-Agnostic Interpretability for Student Depression Screening: Integrating XAI and Multi-Criteria Model Selection. Cureus J Comput Sci 3 : es44389-026-00098-8. DOI https://doi.org/10.7759/s44389-026-00098-8

Feature	Mean (Severe)	Mean (Non-Severe)	*p*-Value	Sig.	Cohen's *d*	Effect	95% CI
BDI Total Score	38.887	13.939	<0.0001	Yes	3.3853	Large	[23.551, 26.347]
CES-D Total Score	15.225	5.204	<0.0001	Yes	2.2495	Large	[9.100, 10.943]
UCLA Loneliness	20.042	15.958	<0.0001	Yes	0.8218	Large	[3.146, 5.022]
Financial Pressure	0.803	0.662	0.0009	Yes	0.3214	Small	[0.059, 0.223]
Social Media Usage	1.542	1.380	0.1184	No	0.1567	Small	[-0.041, 0.365]
Sleep Category	0.761	0.760	0.9908	No	0.0012	Small	[-0.133, 0.135]
Age	22.261	22.291	0.8592	No	-0.018	Small	[-0.360, 0.300]

TABLE 4: Statistical comparison of key features between Severe and Non-Severe students (Welch's t-test).

BDI, Beck Depression Inventory; CES-D, Center for Epidemiologic Studies Depression Scale; UCLA, University of California Los Angeles Loneliness Scale

Explainability analysis

Explainability is analyzed through three interrelated dimensions: global feature significance, directional feature impacts, and localized case-level reasoning, underpinned by quantitative assessments of stability and fidelity to ensure reliability beyond mere visual interpretation. Global analyses show that four psychometric composites, BDI_Total, UCLA_Total, CESD_Total, and PHQ_Total; are the most important for predicting severe risk. Only 4 out of 27 features exceed the attribution threshold ($|\phi| > 0.01$), indicating substantial explanatory sparsity.

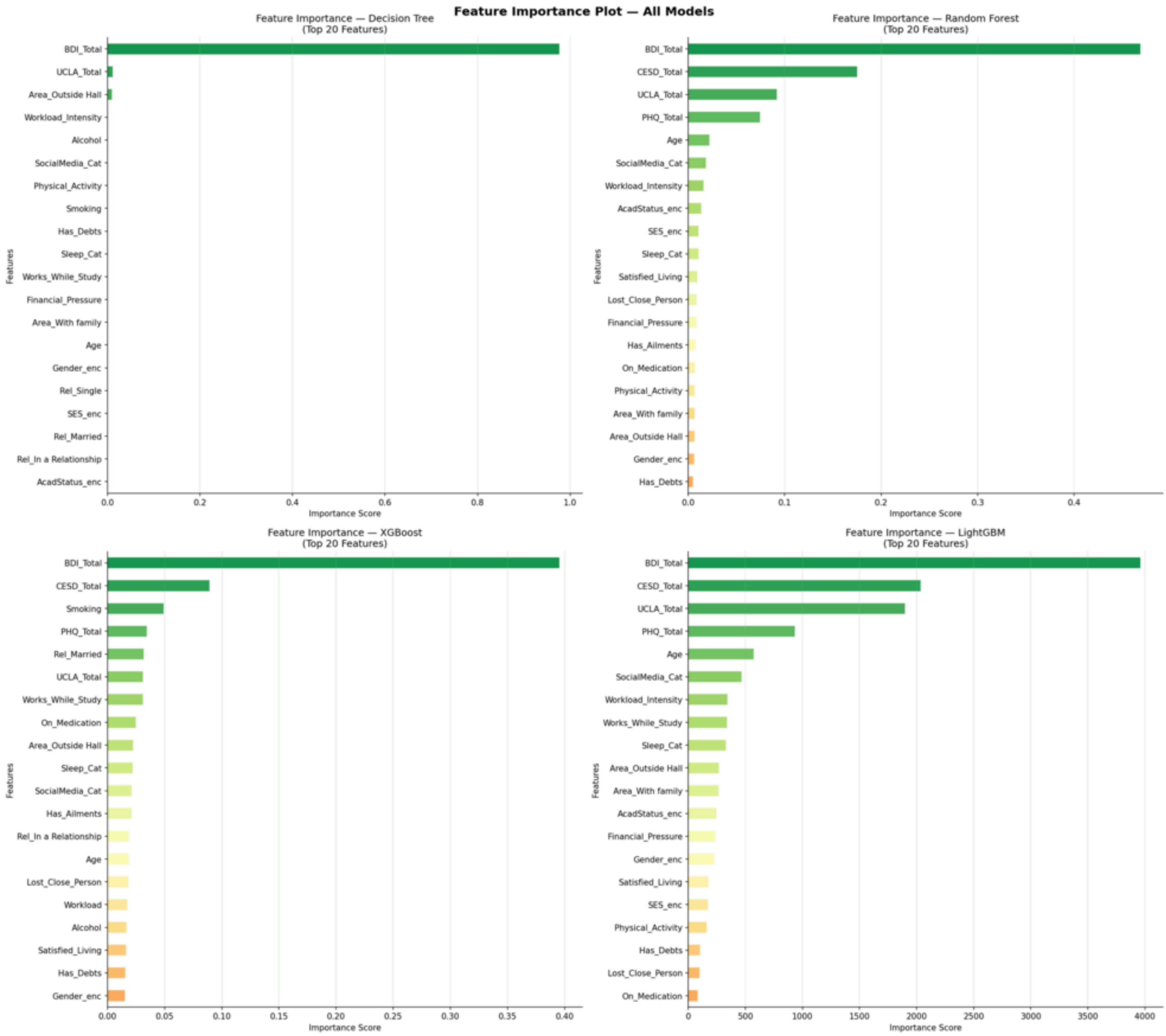

FIGURE 8: Comparative global feature importance across models, highlighting the dominance of psychometric composites (BDI, CES-D, PHQ, UCLA) and the comparatively smaller contribution of demographic/lifestyle variables.

BDI, Beck Depression Inventory; CES-D, Center for Epidemiologic Studies Depression Scale; PHQ, Patient Health Questionnaire; UCLA, University of California Los Angeles Loneliness Scale

Cross-model comparison of global importance indicates that the observed hierarchy is algorithm-robust rather than model-specific. In particular, the dominance pattern visible in Figure 8 is replicated across distinct learners, with psychometric totals consistently ranking at the top, reducing concern that one architecture may have overfit idiosyncratic feature interactions in the training split. It also strengthens the external plausibility of the predictor order by aligning with established clinical interpretations of depression scales. Consequently, the global importance profile can be interpreted as a substantive signal rather than a methodological artifact.

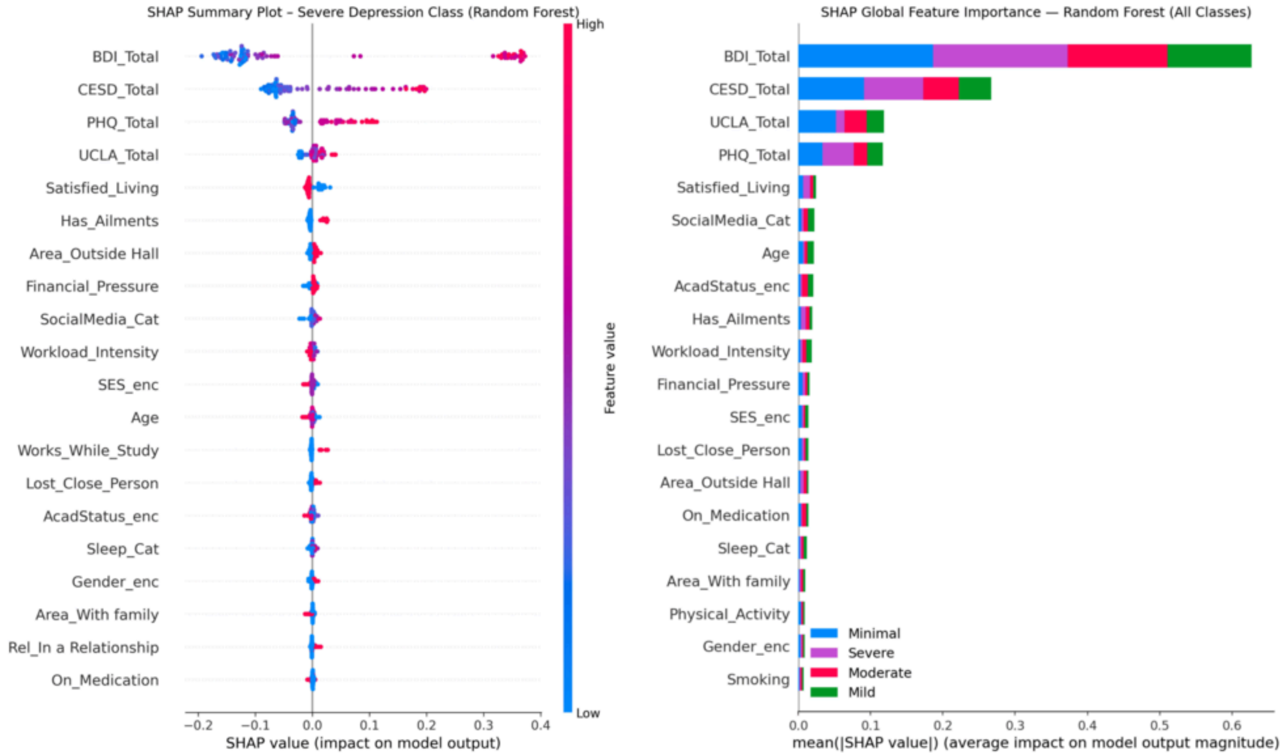

FIGURE 9: SHAP beeswarm summary for global interpretability, showing per-instance attribution distributions for top features.

Color encodes feature value, and horizontal spread encodes magnitude and direction of impact on prediction, enabling simultaneous inspection of importance and effect polarity. SHAP, SHapley Additive exPlanations; BDI, Beck Depression Inventory; CES-D, Center for Epidemiologic Studies Depression Scale; PHQ, Patient Health Questionnaire; UCLA, University of California Los Angeles Loneliness Scale

Importance ranking alone does not show whether a feature increases or decreases risk across individuals. Figure *9* provides that directional information, where color-value encoding and horizontal spread jointly reveal effect polarity and magnitude dispersion. High values of BDI, CES-D, PHQ, and UCLA systematically move outputs toward higher-severity predictions, while lower values exert a protective influence. Most lower-ranked variables remain tightly concentrated around zero attribution, indicating limited marginal contribution for most cases. This distributional behavior further supports a parsimonious interpretation of model reasoning.

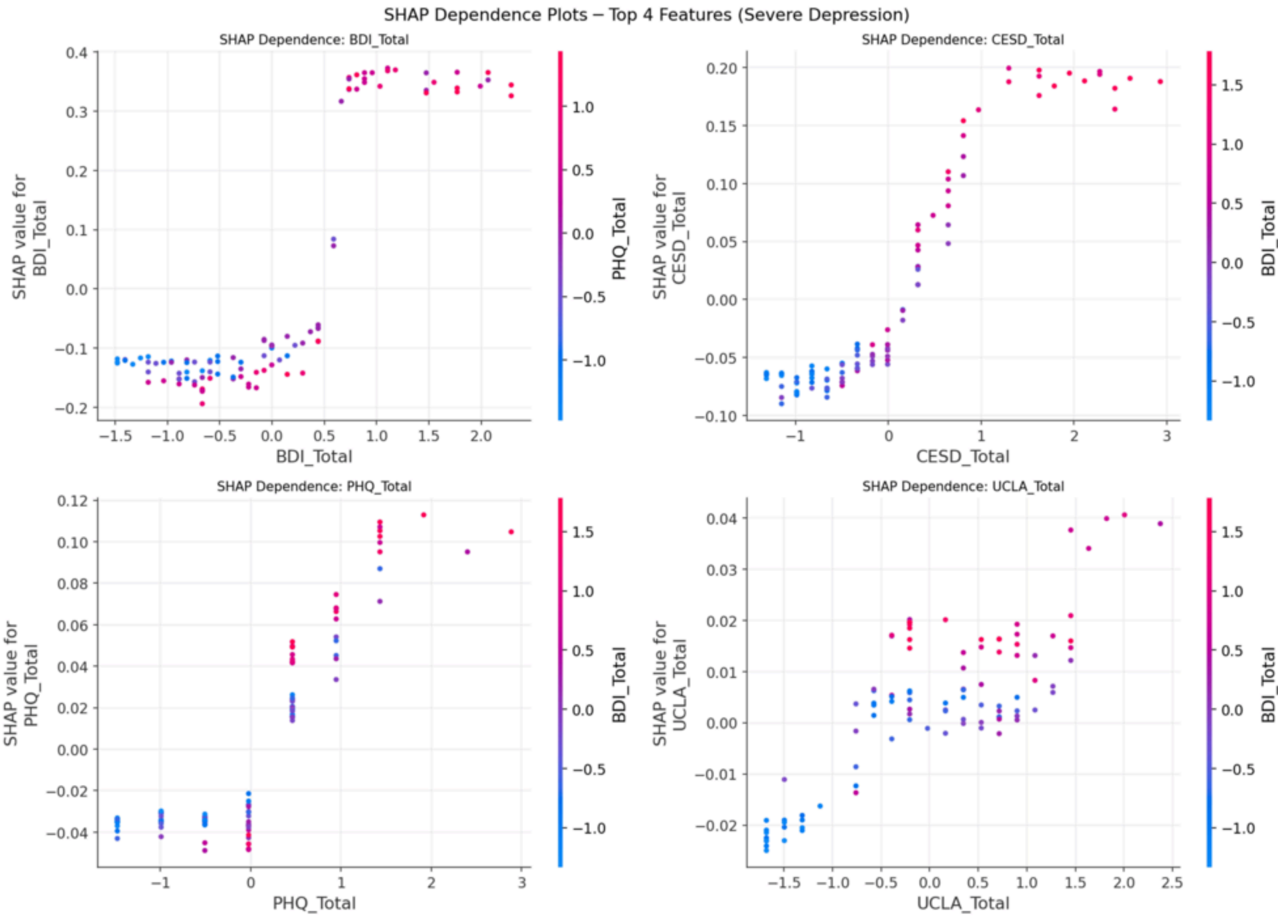

FIGURE 10: SHAP dependence plots for the severe class, illustrating how feature value changes (and interactions) alter attribution magnitude.

The panels reveal threshold-like behavior for key psychometric predictors and comparatively smaller, noisier effects for secondary variables. SHAP, SHapley Additive exPlanations; BDI, Beck Depression Inventory; CES-D, Center for Epidemiologic Studies Depression Scale; PHQ, Patient Health Questionnaire; UCLA, University of California Los Angeles Loneliness Scale

The dependence analysis in Figure *10* shows that BDI and PHQ have transitions that look like thresholds, which is what happens in real life when assigning severe risk. CES-D and UCLA show smoother increases, but both still show positive trends in the high-risk range. Interaction coloring indicates that a heightened BDI amplifies the marginal effects of other psychometric variables, thereby solidifying its position as a pivotal anchor in class-3 inference. This behavior aligns with the documented monotonicity diagnostics and enhances confidence in the directional coherence of explanations.

Explanation stability was evaluated by computing SHAP ranking vectors on five random subsamples of X_{test} (60% each) and measuring pairwise Spearman correlations. The mean correlation was $\rho_s = 0.9741 \pm 0.0091$ (min: 0.9542, max: 0.9884), well above the commonly used stability threshold of 0.90, indicating robustness to sampling variation. Faithfulness for the Severe class was $r_f = 1.0000$, showing exact reconstruction of model logits by $\sum_j \phi_j$, consistent with SHAP efficiency. The perturbation-based infidelity score was low (0.001927 ± 0.003018), indicating limited degradation in explanatory quality under small perturbations.

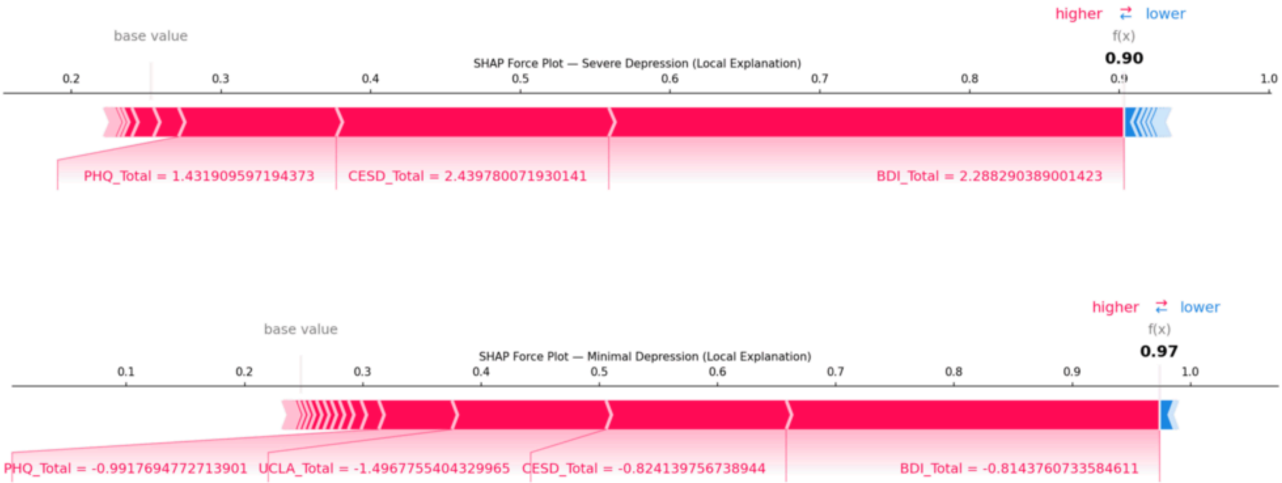

FIGURE 11: SHAP force-plot explanation for representative instances, showing how feature-level positive and negative contributions move predictions from the baseline expectation to class-specific outputs.

Arrow magnitude reflects attribution strength for each feature. SHAP, SHapley Additive exPlanations; BDI, Beck Depression Inventory; CES-D, Center for Epidemiologic Studies Depression Scale; PHQ, Patient Health Questionnaire; UCLA, University of California Los Angeles Loneliness Scale

By showing how each feature moves the output away from the baseline expectation, the local decomposition in Figure 11 makes the evidential direction clear. For severe profiles, elevated psychometric loads generate significant positive shifts towards class-3 assignment, while low-scale and low-loneliness profiles draw outputs towards minimal severity. The additive structure also makes it easier to conduct a post hoc audit by showing whether a decision was based on clinically important variables, making the explanation format good for reviewing and justifying at the individual level.

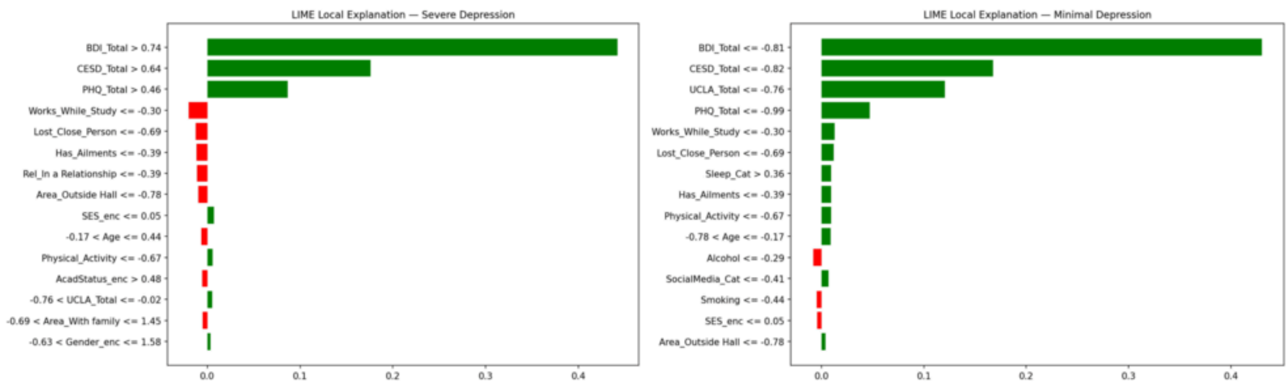

FIGURE 12: LIME local surrogate explanations contrasting representative Severe and Minimal predictions.

Positive and negative local weights indicate which thresholded feature conditions support or oppose each class assignment in the neighborhood of the instance. LIME, Local Interpretable Model-Agnostic Explanations; BDI, Beck Depression Inventory; CES-D, Center for Epidemiologic Studies Depression Scale; PHQ, Patient Health Questionnaire; UCLA, University of California Los Angeles Loneliness Scale

The surrogate-based view in Figure 12 gives sparse, threshold-style rules that are easier to say out loud than full attribution vectors. High BDI/CES-D/PHQ scores mostly support strong predictions, while low psychometric and loneliness levels mostly support weak predictions. This directionality is similar to the SHAP local and global findings, indicating that the methods are consistent rather than contradictory. This convergence increases confidence that the reported local rationale is stable and clinically interpretable.

Model	Train time (s)	Inference time (s)	Model complexity
LightGBM	0.3788	0.0110	300 trees
Decision Tree	0.0057	0.0013	25 nodes
Random Forest	0.7481	0.0987	300 trees
XGBoost	0.3408	0.0051	300 trees

TABLE 5: Computational performance metrics on each models.

Table 5 shows the time it takes to train, the time it takes to make predictions, and the complexity of the model. Decision Tree is the most efficient model for computation. It trains in 5.7 ms and makes decisions in 1.3 ms per batch, and it only has 25 internal nodes in its tree structure. Logistic Regression is also fast (training: 45.3 ms, inference: 1.8 ms) because it uses a linear parameterization (108 parameters for a 4-class, 27-feature multinomial model). Random Forest, XGBoost, and LightGBM all use 300 estimators, but they take different amounts of time to train (748 ms, 341 ms, and 379 ms, respectively). Inference latencies are all below 11 ms per batch, which is well within the requirements for real-time screening.

Comparative analysis discussion

This section puts the real-world results in the context of earlier studies on student mental health, explainable machine learning, and the limits of deployment in resource-constrained settings. By evaluating not only predictive performance but also interpretability, robustness, and practical utility by combining accuracy, calibration, fairness, statistical effect sizes, and attribution stability. Table 6 situates the current study within the extensive literature, emphasizing both alignment with established results and significant methodological progress.

References	Domain	Previous key findings	Present study
[3,22]	Depression Prevalence	33.2% (BD); 35.4% Mild, 13.4% Severe (global)	27.0% Moderate, 15.6% Severe
	Anxiety Prevalence	29.0% (BD); 40.2% Mild, 16.8% Severe (global)	21.6% moderate, 12.6% Severe
[8]	Sleep & Mental Health	Strongest predictor ($\beta = 3.341$)	Not independently significant ($p = 0.991$); interacts with composites
[4,18]	Academic/Financial Stress	Significant negative correlation with well-being; OR = 3.11	Financial pressure significant ($p = 0.001$, $d = 0.32$)
[18]	Machine Learning Accuracy	Worsening symptom prediction: balanced accuracy = 0.49	Best accuracy 96.0% (DT, LightGBM); XGBoost ECE = 0.0185
[6,17]	Fairness in ML	Limited prior work; female and LGBTQ+ students at higher risk	Random Forest: EOD = 0.000; DPD = 0.219–0.246
[31]	Explainability (XAI)	No quantitative stability/fidelity metrics reported	SHAP: ρ's = 0.974, rf = 1.000, infidelity = 0.0019
[12,13]	Treatment Gap	Only 8.5–24% of affected students receive care	Not assessed (screening-focused)
[24]	Longitudinal Evidence	60% persistence at 2 years	Cross-sectional design

TABLE 6: Comparison of key findings between the present study and existing literature.

BD, Bipolar Disorder; XAI, Explainable Artificial Intelligence; OR, Odd Ratio; DT, Decision Trees; ECE, Estimated Calibration Error; EOD, Equal Opportunity Difference; DPD, Demographic Parity Difference; SHAP, SHapley Additive exPlanations; LGBTQ+, Lesbian, Gay, Bisexual Transgender, or Queer plus other sexual orientations and gender identities

The findings demonstrate that composite psychometric scores, specifically BDI-II, CES-D, PHQ, and UCLA Loneliness, are the primary indicators of depression severity. It corroborates previous findings indicating that validated instruments surpass singular behavioral or demographic characteristics [18,31]. However, the findings expand this literature by showing that a small set of composite features accounts for most of the predictive signal. At the same time, lifestyle variables mostly do so through interaction effects, which implies that the principal advantage of the machine learning framework is not in identifying new predictors, but in successfully amalgamating validated metrics into a cohesive, interpretable, and probabilistically calibrated screening model. The Decision Tree has a simple, easy-to-understand structure and an accuracy of 96.00%, which supports earlier suggestions that it should be used in clinical screening settings [15]. XGBoost has better probability calibration (ECE = 0.0185), which is important for risk stratification and resource allocation in real-world scenarios. In resource-limited settings like Bangladesh, accurately calibrated probabilities are crucial for reducing the misallocation of scarce mental health services [3,11]. So, depending on the deployment goals, model selection should strike a balance between interpretability and calibration.

Several variables identified as risk factors in prior studies, including sleep, social media use, and age, are not independently significant in this dataset, highlighting the distinction between univariate association and multivariate predictive contribution [3,19,29]. Financial pressure remains significant but with a modest effect size, suggesting indirect influence mediated through psychometric constructs [18]. Fairness analysis reveals zero Equal Opportunity Difference for Random Forest, indicating equitable true-positive rates across genders, although moderate Demographic Parity Differences may reflect underlying prevalence variations [14,17,22]. Finally, explainability results demonstrate high stability, faithfulness, and low infidelity, addressing limitations in prior work lacking quantitative validation of explanations [18,21,31]. Notably, psychometric features exhibit consistent monotonic relationships with predicted severity, thereby enhancing clinical interpretability, whereas behavioral features show more context-dependent effects.

Conclusions

This research proposes an XAI model for predicting depression severity among university students in Bangladesh by employing various tree-based machine learning algorithms and incorporating SHAP and LIME interpretability methods. To support strong and transparent model evaluation, a multi-criteria evaluation strategy was used that included accuracy, calibration, fairness, stability, and explainability. Based on experimental results, the Decision Tree and LightGBM classifiers are the most accurate, and XGBoost has the best probabilistic calibration. The results of the SHAP analysis indicate that composite psychometric scores (BDI and CES-D) are the most influential predictors; LIME provides complementary local-level explanations that enhance interpretability and trustworthiness. The results showed that explainable, multi-objective machine learning models can provide dependable, trustworthy guidance in university mental health screening. The study, however, has limitations, including the use of only one cross-sectional data set, which may not be generalizable to other populations. Although some limitations are noted, results from several evaluation metrics and explanation techniques are consistent, strengthening the validity of the conclusions and the applicability of interpretable machine learning systems for scalable mental health assessment.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Md. Abid Hasan Rafi, Mohima Binte Rasel, Mst. Fatematuj Johora, Pankaj Bhowmik

Acquisition, analysis, or interpretation of data: Md. Abid Hasan Rafi, Mohima Binte Rasel, Mst. Fatematuj Johora, Pankaj Bhowmik

Drafting of the manuscript: Md. Abid Hasan Rafi, Mohima Binte Rasel, Mst. Fatematuj Johora, Pankaj Bhowmik

Critical review of the manuscript for important intellectual content: Md. Abid Hasan Rafi, Mohima Binte Rasel, Mst. Fatematuj Johora, Pankaj Bhowmik

Supervision: Md. Abid Hasan Rafi, Pankaj Bhowmik

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial**

relationships: All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The data (and/or code) supporting this study are openly available in Mendeley Data at https://data.mendeley.com/datasets/f4z2bfv7vk/1, reference 10.17632/f4z2bfv7vk.1.

Acknowledgements

The authors would like to acknowledge the use of Gemini 3.1 Pro for generating the scientific-grade methodological illustration presented in this work.

References

1. Adams KL, Saunders KE, Keown-Stoneman CDG, Duffy AC: Mental health trajectories in undergraduate students over the first year of university: a longitudinal cohort study. BMJ Open. 2021, 11:1-11. 10.1136/bmjopen-2020-047393
2. Ahluwalia M, Shillington KJ, Irwin JD: The relationship between resilience and mental health of undergraduate students: a scoping review. Journal of American College Health. 2023, 73:1266-1279. 10.1080/07448481.2023.2252925
3. Al-Amin M, Rinky F, Nizamul HBM, Yeasmin R, Akter T, Hoque N, Reza S: Factors influencing mental health outcomes among university students: a cross-sectional study in Bangladesh. BMJ Open. 2025, 15:1-16. 10.1136/bmjopen-2024-097745
4. Barbayannis G, Bandari M, Zheng X, Baquerizo H, Pecor KW, Ming X: Academic stress and mental well-being in college students: correlations, affected groups, and COVID-19. Frontiers in Psychology. 2022, 13:1-10. 10.3389/fpsyg.2022.886344
5. Botella-Juan L, Arias-De la Torre J, Amezcua-Prieto C, et al.: Evolution of the risk of problematic alcohol use in university students: a longitudinal study based on the uniHcos cohort data. International Journal of Mental Health and Addiction. 2025, 1-19. 10.1007/s11469-025-01512-3
6. Campbell F, Blank L, Cantrell A, Baxter S, Blackmore C, Dixon J, Goyder E: Factors that influence mental health of university and college students in the UK: a systematic review. BMC Public Health. 2022, 22:1-22. 10.1186/s12889-022-13943-x
7. Emmerton RW, Camilleri C, Sammut S: Continued deterioration in university student mental health: Inevitable decline or skirting around the deeper problems?. Journal of Affective Disorders Reports. 2024, 15:100691. 10.1016/j.jadr.2023.100691
8. Erol S, Gur K, Karaca S, Çalık KB, Uzuner A, Apaydın Kaya Ç: Risk factors affecting the mental health of first-year university students on a health sciences campus and related factors. The Journal of Mental Health Training, Education and Practice. 2022, 18:146-157. 10.1108/jmhtep-03-2022-0015
9. Hamilton J, Welham A, Morgan G, Jones C: Exploring the prevalence of childhood adversity among university students in the United Kingdom: a systematic review and meta-analysis. PLOS One. 2024, 19:1-24. 10.1371/journal.pone.0308038
10. A Dataset on the Mental Health of Undergraduate Students in Bangladesh. (2024). Accessed: March 28, 2026: https://data.mendeley.com/datasets/f4z2bfv7vk/1.
11. Hoque R: Major mental health problems of undergraduate students in a private University of Dhaka, Bangladesh. European Psychiatry. 2015, 30:1880. 10.1016/s0924-9338(15)31442-5
12. Hunt J, Eisenberg D: Mental health problems and help-seeking behavior among college students. Journal of Adolescent Health. 2010, 46:3-10. 10.1016/j.jadohealth.2009.08.008
13. King N, Pickett W, McNevin SH, et al.: Mental health need of students at entry to university: baseline findings from the U-Flourish Student Well-Being and Academic Success Study. Early Intervention in Psychiatry. 2020, 15:286-295. 10.1111/eip.12939
14. Lee J, Jeong HJ, Kim S: Stress, anxiety, and depression among undergraduate students during the COVID-19 pandemic and their use of mental health services. Innovative Higher Education. 2021, 46:519-538. 10.1007/s10755-021-09552-y
15. Limone P, Toto GA: Factors that predispose undergraduates to mental issues: a cumulative literature review for future research perspectives. Frontiers in Public Health. 2022, 10:1-12. 10.3389/fpubh.2022.831349

16. Lipson SK, Diaz Y, Davis J, Eisenberg D: Mental health among first-generation college students: findings from the national healthy minds study, 2018-2021. Cogent Mental Health. 2023, 2:1-15. 10.1080/28324765.2023.2220358

17. Liu Y, Frazier PA, Porta CM, Lust K: Mental health of US undergraduate and graduate students before and during the COVID-19 pandemic: differences across sociodemographic groups. Psychiatry Research. 2022, 309:1-7. 10.1016/j.psychres.2022.114428

18. Meda N, Pardini S, Rigobello P, Visioli F, Novara C: Frequency and machine learning predictors of severe depressive symptoms and suicidal ideation among university students. Epidemiology and Psychiatric Sciences. 2023, 32:1-13. 10.1017/s2045796023000550

19. Miles R, Rabin L, Krishnan A, Grandoit E, Kloskowski K: Mental health literacy in a diverse sample of undergraduate students: demographic, psychological, and academic correlates. BMC Public Health. 2020, 20:1-13. 10.1186/s12889-020-09696-0

20. Milojevich HM, Lukowski AF: Sleep and mental health in undergraduate students with generally healthy sleep habits. PLOS One. 2016, 11:1-14. 10.1371/journal.pone.0156372

21. Ou J, Wang K, Zuo M, Chen D, Luo H: Factors influencing the mental health of first-year college students: evidence from digital records of daily behaviors. Behavioral Sciences. 2025, 15:1-23. 10.3390/bs15050618

22. Paiva U, Cortese S, Flor M, et al.: Prevalence of mental disorder symptoms among university students: an umbrella review. Neuroscience & Biobehavioral Reviews. 2025, 175:1-10. 10.1016/j.neubiorev.2025.106244

23. Park SY, Andalibi N, Zou Y, Ambulkar S, Huh-Yoo J: Understanding students' mental well-being challenges on a university campus: interview study. JMIR Formative Research. 2020, 4:e15962. 10.2196/15962

24. Pedrelli P, Nyer M, Yeung A, Zulauf C, Wilens T: College students: mental health problems and treatment considerations. Academic Psychiatry. 2014, 39:503-511. 10.1007/s40596-014-0205-9

25. Read A, Lutgens D, Malla A: A descriptive overview of mental health services offered in post-secondary educational institutions across Canada. The Canadian Journal of Psychiatry. 2022, 68:101-108. 10.1177/07067437221128168

26. Rodríguez-Romo G, Acebes-Sánchez J, García-Merino S, Garrido-Muñoz M, Blanco-García C, Diez-Vega I: Physical activity and mental health in undergraduate students. International Journal of Environmental Research and Public Health. 2022, 20:1-12. 10.3390/ijerph20010195

27. Salimi N, Gere B, Talley W, Irioogbe B: College students mental health challenges: concerns and considerations in the COVID-19 pandemic. Journal of College Student Psychotherapy. 2021, 37:39-51. 10.1080/87568225.2021.1890298

28. Valdés JM, Díaz FJ, Christiansen PM, et al.: Mental health and related factors among undergraduate students during SARS-CoV-2 pandemic: a cross-sectional study. Frontiers in Psychiatry. 2022, 13:1-20. 10.3389/fpsyt.2022.833263

29. Wyatt T, Oswalt SB: Comparing mental health issues among undergraduate and graduate students. American Journal of Health Education. 2013, 44:96-107. 10.1080/19325037.2013.764248

30. Xie H, Kang B: Mental health of nursing students: a bibliometric review based on CiteSpace visual analysis. Journal of Nursing Management. 2025, 2025:1-11. 10.1155/jonm/2169094

31. Zhang J, Peng C, Chen C: Mental health and academic performance of college students: knowledge in the field of mental health, self-control, and learning in college. Acta Psychologica. 2024, 248:1-10. 10.1016/j.actpsy.2024.104351

32. Kadirvelu B, Bellido BT, Freccero A, Di Simplico M, Nicholls D, Faisal AA: Digital phenotyping for adolescent mental health: feasibility study using machine learning to predict mental health risk from active and passive smartphone data. Journal of Medical Internet Research. 2026, 28:1-27. 10.2196/72501

33. Othman NA, Rosdi M : Machine learning approaches to workplace mental health: predicting treatment-seeking behavior using the OSMI dataset. Tech-E. 2026, 9:166-179. 10.31253/te.v9i2.4305

34. Biswas N, Maiti S, Kundu S, Biswas S : Advancing mental health insights through machine learning on EEG data. Integrative Machine Learning and Optimization Algorithms for Disease Prediction. Muniasamy A (ed): IGI Global Scientific Publishing, 2026. 315-338. 10.4018/979-8-3373-1087-9.ch011

35. Lamba K, Rani S, Shabaz M: Explainable machine learning for mental health prediction from social media behavior: a nested cross-validation study with SHAP and LIME interpretability. Discover Mental Health. 2026, 6:1-30. 10.1007/s44192-026-00373-z

36. Chang Y, Feng Y, Sun J, Ai J, Li C, Zhou K, Zhang K: SridBench: benchmark of scientific research illustration drawing of image generation model. arXiv. 2025, 1-13. 10.48550/arXiv.2505.22126

37. AbidHasanRafi/ML-based-Mental-Health-XAI. (2026). Accessed: May 9, 2026: https://github.com/AbidHasanRafi/ML-based-Mental-Health-XAI.

How to cite this article:
Rasel M, Rafi M, Johora M, et al. (June 03, 2026) Model-Agnostic Interpretability for Student Depression Screening: Integrating XAI and Multi-Criteria Model Selection. Cureus J Comput Sci 3 : es44389-026-00098-8. DOI https://doi.org/10.7759/s44389-026-00098-8