RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt26513

Early Depression Risk Screening using Leakage-Aware Ensemble Learning with Probability Calibration

Mohitsinh Parmar [a, *], Sohil D. Pandya [b]

[a] Department of Computer Science and Applications, CMPICA, Charotar University of Science and Technology (CHARUSAT), Gujarat, India.

* Corresponding Author Email: mohitsinhparmar.mca@charusata.ac.in

DOI: https://doi.org/10.54392/irjmt26513

Received: 30-01-2026; Revised: 22-08-2026; Accepted: 15-09-2026; Published: 22-09-2026

Abstract: Depression among university students is an important mental health concern, particularly in technical and engineering education environments characterized by academic workload, competition, and financial stress. Although machine learning approaches are increasingly used in student mental health research, many previous studies rely on leakage-prone or post-outcome variables, limited validation protocols, and discrimination-focused evaluation, which may reduce their suitability for realistic early screening settings. This study presents a leakage-aware and calibration-aware ensemble learning framework for early depression risk screening among technical education students by excluding leakage-prone predictors during model development. Random Forest, Gradient Boosting, and stacking-based ensemble models were evaluated using nested stratified cross-validation, pooled out-of-fold prediction analysis, isotonic probability calibration, threshold sensitivity analysis, subgroup robustness assessment, and SHAP-based interpretability analysis. Experiments conducted on a publicly available dataset filtered to technical degree programs (N = 7,807) showed stable cross-validated performance, with the calibrated stacking ensemble achieving ROC–AUC = 0.8718 ± 0.0081, PR–AUC = 0.8888 ± 0.0078, and recall = 0.8408 ± 0.0100. Statistical testing showed no significant performance difference between stacking and calibrated Gradient Boosting, suggesting that simpler calibrated models may provide comparable screening performance in this dataset. Leakage-ablation analysis showed that inclusion of the post-outcome suicidal-thoughts variable increased pooled out-of-fold ROC–AUC from 0.8701 to 0.9224, highlighting the importance of leakage-aware feature selection for realistic evaluation. A reduced-feature screening model using six early-available predictors also maintained competitive performance (ROC–AUC = 0.8616), supporting the feasibility of lightweight institutional screening. SHAP stability analysis demonstrated consistent feature rankings across validation folds, with academic pressure, financial stress, work/study hours, and sleep duration identified as influential predictors. Overall, the proposed framework provides an interpretable and methodologically transparent approach for depression risk prioritization in educational settings while reducing performance inflation associated with leakage-prone features.

Keywords: Depression Risk Screening, Ensemble Learning, Leakage-Aware Modeling, Probability Calibration, Student Mental Health Analytics.

1. Introduction

Depression is a prevalent disabling condition among young adults and is associated with diminished quality of life and substantial academic, social, and economic consequences worldwide [1]. The academic burden, financial pressures and psychosocial adjustment are important risk factors for students at university. Meta-analytic evidence indicates that depressive symptoms affect approximately 25–30% of university students across diverse educational contexts [2, 3]. Students enrolled in technically demanding disciplines such as engineering and computer science may experience additional stress due to intensive coursework, continuous evaluation, and employability-related expectations [4]. Further, students are likely to suffer from psychological distress due to financial stress, which has been shown to be a strong predictor of student distress [5].

Due to their ability to capture complex nonlinear relationships in student behaviour and learning patterns, machine learning approaches have received increasing attention for student mental-health screening [6-8]. However, many existing studies suffer from methodological limitations that reduce their applicability for early screening. The inclusion of leakage-prone variables can artificially inflate predictive performance and reduce the realism of deployment-oriented

evaluation. Several studies also rely on a single train–test split and report only discrimination metrics (accuracy and ROC–AUC) without measuring calibration, uncertainty, or threshold sensitivity.

In mental health screening, false-negative predictions can have serious consequences. Healthcare analytics frameworks should prioritize recall-oriented evaluation of predictive systems, the temporal plausibility of screening variables, and operational constraints on decision-making rather than maximizing classification accuracy [9, 10]. Additionally, modern ML models often produce poorly calibrated probabilities despite strong discrimination performance, limiting their reliability for deployment-oriented risk prioritization [11, 12].

In this study, leakage-aware modeling refers to a screening-oriented feature-restriction strategy that excludes variables unavailable before intervention or closely associated with downstream symptom manifestations. The proposed framework does not perform causal inference, counterfactual estimation, or causal identification. Instead, it seeks to preserve deployment realism through temporally plausible feature selection. Explainable artificial intelligence techniques are also used to support interpretation in a high-stakes screening context. Although feature-importance analysis has been reported in earlier studies, relatively few investigations have examined the stability of explanations across validation folds or model fits [13, 14].

Accordingly, this study aims to develop and evaluate a leakage-aware, calibration-aware, and interpretable ensemble-learning framework for early depression-risk screening among technical education students. The study focuses on deployment-oriented evaluation by integrating temporally plausible feature selection, pooled out-of-fold prediction, probability calibration, threshold-sensitivity analysis, subgroup robustness assessment, and SHAP-based interpretability.

1.1 Background and Related Studies

The growing availability of educational, behavioural, and psychosocial data has encouraged the use of machine learning for student mental-health screening. Earlier research has examined depression risk using statistical methods as well as supervised learning models. However, many studies treat university students as a broadly homogeneous population and give limited attention to the distinct stress patterns associated with particular academic disciplines. Engineering, computing and other technical disciplines can involve ongoing workload, regular assessment, work-related stress and financial issues, which can become a burden to students' mental health [15].

Socioeconomic and financial factors are widely discussed in epidemiological research, but they are less consistently incorporated into machine learning-based screening systems [16]. As a result, some existing models provide general risk classifications without fully considering whether the selected predictors are suitable for early institutional screening. Broader reviews of machine learning in mental health have also noted considerable variation in data sources, outcome definitions, validation practices, and reporting quality [6].

Recent student-focused studies have applied a range of machine learning methods to mental-health screening. Bhattacharjee et al. used machine learning models to predict depression among public-university students [17], while Sonjaya *et al.* applied a Naive Bayes classifier to a publicly available student depression dataset [18]. Xu and Zhang used boosting and linguistic features to classify mental-health conditions in online-learning environments [19]. A family-informed study combined individual and parental factors using sparse Logistic Regression, Support Vector Machine, and Random Forest models [20]. Meda *et al*. adopted a longitudinal design to examine changes in severe depressive symptoms and suicidal ideation over six months, although identifying newly developing or worsening cases remained difficult [21]. Other studies have investigated stacking ensembles with oversampling [22] and tree-based models for the joint prediction of anxiety, depression, and insomnia [23]. Zhang *et al.* evaluated a severe mental-distress model using an independent external cohort, probability-reliability measures, and SHAP-based interpretation [24]. Hasan *et al*. compared several machine learning algorithms and showed that apparently acceptable accuracy may coexist with relatively weak discrimination, emphasizing the need to interpret accuracy together with ROC–AUC and class-specific measures [25]. Collectively, these studies show increasing interest in ensemble learning, longitudinal assessment, external validation, and explainability, although such practices are not yet consistently adopted across student mental-health research.

Ensemble and tree-based methods, including Random Forest and Gradient Boosting, are frequently used because they can capture nonlinear relations among academic, demographic, lifestyle, and psychosocial variables [26]. Nevertheless, several published studies rely on a single train–test split, report mainly discrimination measures, or do not explicitly examine whether predictors would be available before the outcome or intervention. Probability calibration, threshold-sensitive evaluation, and uncertainty analysis are also less commonly reported, although they are important when predicted risks are intended to guide screening priorities [27, 28].

Table 1. Comparison of selected recent machine learning studies on student mental-health screening.

Study	Cohort / Outcome	Model and validation	Deployment-related element	Main limitation
Gil *et al*. [20]	Korean students; depression risk	LR, SVM, RF; holdout test	Family predictors and feature importance	Small sample; no external validation
Meda *et al*. [21]	Italian students; depression and suicidal-ideation trajectories	RF; six-month follow-up	Longitudinal prediction	Weak detection of worsening cases
Daza Vergaray *et al*. [22]	Engineering students; depression levels	Stacking with oversampling; cross-validation	Ensemble classification	Small, single-discipline sample
Chowdhury *et al*. [23]	Bangladeshi students; anxiety, depression, and insomnia	DT, RF, XGBoost; train–test evaluation	Multi-outcome prediction	Cross-sectional; no external validation
Zhang *et al.* [24]	Chinese students; severe mental distress	Multiple ML models; external validation	Calibration and SHAP analysis	Composite outcome
Hasan *et al.* [25]	Bangladeshi students; depression, anxiety, and stress	Seven ML models; cross-validation	Training-only SMOTE and SHAP	Weak discriminatory performance
Present study	Technical education students; depression-risk label	RF, GB, stacking; nested CV and pooled OOF	Leakage control, calibration, threshold and SHAP stability analysis	Single-dataset evaluation

Interpretability and ethical transparency are further concerns in student mental-health analytics [29, 30]. A recent systematic review of machine learning for mental-health prediction identified continuing challenges related to dataset heterogeneity, model transparency, and generalizability [31]. An ethics-centred review of student mental-health studies found that privacy, fairness, consent, and ethical reporting were often insufficiently discussed [32]. These concerns are particularly relevant when models use sensitive behavioural or psychosocial information. Screening tools should therefore support human review rather than replace professional assessment or institutional decision-making [30].

Table 1 presents selected recent empirical studies that are directly comparable with the present work in terms of student population, predictive modelling, validation strategy, and deployment-oriented evaluation.

Collectively, the reviewed studies show increasing use of ensemble learning, external validation, longitudinal assessment, and explainable artificial intelligence in student mental-health screening. However, leakage control, probability calibration, threshold-sensitive evaluation, and stability of model explanations remain inconsistently examined. The present study addresses this specific methodological gap by combining these components within a deployment-oriented evaluation framework.

2. Method

This section describes the dataset, preprocessing procedures, model architecture, and evaluation protocol used to develop a leakage-aware and calibration-aware framework for early depression risk screening among university students. The overall workflow consisted of: (i) filtering the dataset to technical education students, (ii) preprocessing demographic, academic, and lifestyle variables, (iii) applying leakage-aware feature selection, (iv) training baseline and ensemble models, (v) calibrating predicted probabilities, and (vi) evaluating performance using stratified cross-validation with discrimination, calibration, and robustness metrics.

2.1 Dataset Description

The study used a publicly available student mental health dataset obtained from Kaggle [33], containing demographic, academic, lifestyle, and psychosocial attributes. The original dataset included 27,901 records spanning multiple professions and educational domains. To improve domain specificity, the dataset was filtered to include only individuals identified

as Students and enrolled in technical or engineering-related programs, including B.Tech, M.Tech, BCA, MCA, B.E, M.E, B.Sc, and B.Arch. The final analytical cohort consisted of 7,807 students.

These degree programs were grouped into a broader technical-education category because they share structurally intensive academic environments characterized by sustained workload, continuous evaluation, extended study hours, and employability-related pressure. Although the cohort remains heterogeneous in disciplinary specialization, the selected programs collectively represent academically demanding educational settings associated with elevated psychological burden in prior literature. The relatively broad age distribution (26.97 ± 4.20 years) reflects the inclusion of postgraduate and professional technical programs in addition to undergraduate cohorts. Accordingly, the study emphasizes shared workload-intensive educational structures rather than strict disciplinary uniformity.

The target variable, Depression, was formulated as a binary classification task:

- Class 1: High-risk depression
- Class 0: Low-risk depression

The task is intended for early risk screening rather than clinical diagnosis. The Depression label was obtained directly from the Kaggle dataset, which provides a pre-annotated binary depression indicator. However, the original survey instrument, scoring methodology, and clinical cutoff criteria were not fully documented by the dataset authors. Therefore, the proposed framework should be interpreted as prediction of a dataset-provided depression-risk label rather than formal psychiatric diagnosis. Because the dataset is publicly processed and partially anonymized, details regarding survey administration, recruitment strategy, and label construction remain unavailable. The demographic and class characteristics of the final analytical cohort are summarized in Table 2.

Table 2. Cohort statistics of the technical student population (N = 7,807).

Characteristic	Value
Total samples	7,807
Class distribution	Class 1: 4,373 (56.01%), Class 0: 3,434 (43.99%)
Imbalance ratio	1.27
Gender distribution	Male: 4,455 (57.06%), Female: 3,352 (42.94%)
Age (years)	26.97 ± 4.20
Academic Pressure	3.06 ± 1.39
Financial Stress	3.11 ± 1.44

2.2 Preprocessing and Leakage-Aware Feature Engineering

Preprocessing was performed to ensure model compatibility, minimize target leakage, and preserve the temporal plausibility required for early screening. Missing values for the Financial Stress variable were imputed using the median to reduce sensitivity to extreme observations.

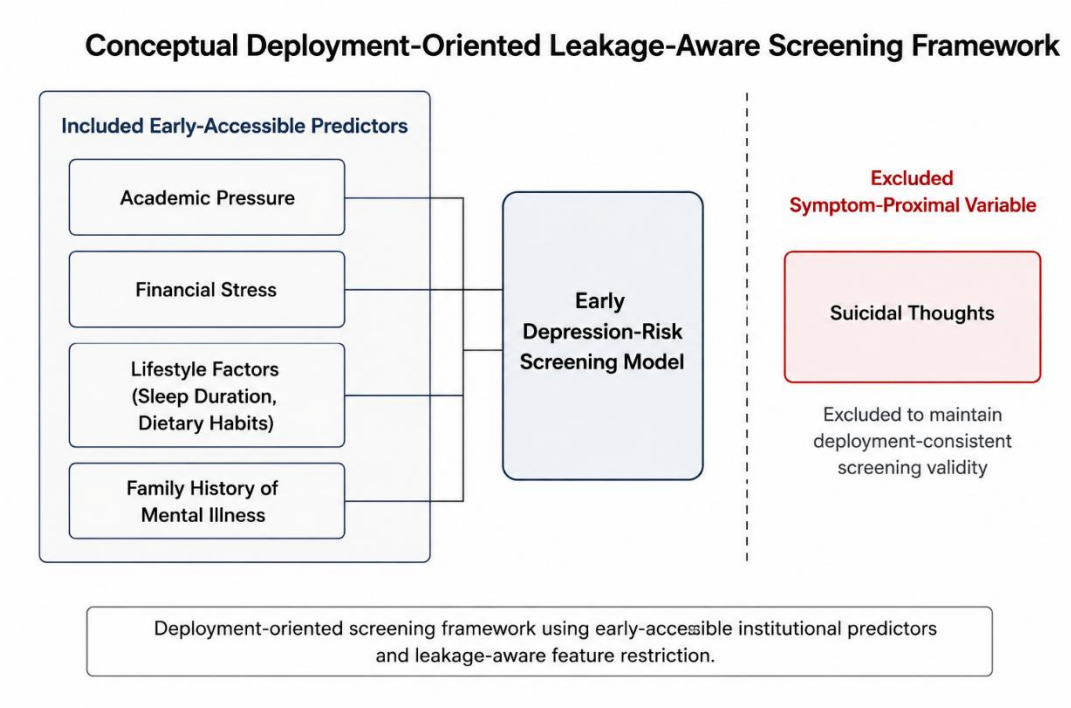


Figure 1. Leakage-aware screening-oriented feature design framework.

Ordinal variables, including Sleep Duration and Dietary Habits, were encoded using ordered mappings. Binary variables, including Gender and Family History of Mental Illness, were label encoded, whereas multiclass categorical variables such as Degree were transformed using one-hot encoding with reference-category removal.

A leakage-aware feature restriction strategy was then applied so that model development relied only on variables that could plausibly be available before intervention or diagnosis. Identifier and location-related variables, including id and City, were removed because they were not considered suitable for institutional screening. Profession was also excluded because it had already been used to filter the dataset to the student population. In addition, the suicidal-thoughts attribute was removed because it represents a symptom-level or downstream indicator that may be closely related to the target outcome.

This feature-selection strategy was intended to prevent the models from learning direct manifestations of depression that may not be appropriate for realistic early-stage screening. The retained variables therefore consisted mainly of academic, financial, demographic, and lifestyle information that could reasonably be collected before risk prioritization. The proposed approach should be interpreted as temporally consistent, screening-oriented feature design rather than as a method for causal identification or counterfactual inference.

As illustrated in Figure 1, only early-accessible academic, financial, demographic, and lifestyle variables were retained for prediction, whereas the suicidal-thoughts variable was excluded to reduce leakage. The retained and removed feature groups are summarized in Table 3.

Table 3. Feature selection strategy (retained vs removed**.**

Category	Description
Retained features	Academic, financial, lifestyle, and demographic variables (e.g., Academic Pressure, Financial Stress, Sleep Duration, CGPA, Work/Study Hours, Family History)
Removed features	id, City, Profession, and suicidal ideation variable (post-outcome proxy)

2.3 Proposed Learning Architecture

A stacking-based ensemble architecture was used to determine whether combining complementary tree-based models could improve prediction robustness while retaining reasonable interpretability. Random Forest and Gradient Boosting were used as the base learners. Random Forest uses bagging to reduce variance and decision trees to capture nonlinear interactions, whereas Gradient Boosting sequentially minimizes prediction error by adding successive weak learners. During the robustness analysis, Gradient Boosting was evaluated both with and without class weighting, while Random Forest was trained using balanced class weights.

Random Forest and Gradient Boosting were used to create the predictions that were combined by a meta-learner: Logistic Regression. During stacking, the meta-learner was trained using cross-validated predictions generated by the base models, rather than predictions from the outer validation fold. This structure has been chosen to minimize overfitting at the ensemble level while maintaining a relatively simple ensemble combination mechanism. The corresponding model configurations are summarized in Table 4.

Table 4. Model configurations

Model	Configuration
Random Forest	n_estimators = 100, max_depth ∈ {5,10}, class_weight = balanced
Gradient Boosting	n_estimators = 100, learning_rate ∈ {0.05,0.1}
Stacking	RF + GB → Logistic Rgression meta-learner, CV = 5

Hyperparameter optimization was done by grid search with five-fold cross validation and stratification. PR-AUC was selected since it balances precision and recall and it can be used for screening oriented evaluation. The search was performed solely on the respective training partitions without taking the outer validation samples into account to avoid their impact on the model selection. The hyperparameter search spaces and selected configurations are reported in Table 5.

Table 5. Hyperparameter search space and selected configurations

Model	Search Space	Best Configuration
Random Forest	n_estimators, max_depth	100, 10
Gradient Boosting	n_estimators, learning_rate	100, 0.1

To ensure reliable probability estimates, isotonic regression was applied after model fitting. Assessment of calibration was done with the Brier score and Expected Calibration Error. ECE was calculated from the pooled out-of-fold predictions using 10 equal-width probability bins. The pre- and post-calibration Brier and ECE values were compared to see if isotonic calibration enhanced the agreement between the predicted risks and the observed frequencies of outcomes. This

analysis was included because discrimination alone does not indicate whether a predicted probability is a reliable estimate of outcome likelihood.

2.4 Evaluation Protocol

All experiments were implemented in Python 3.12 using scikit-learn, NumPy, pandas, matplotlib, seaborn, and SHAP. The experimental pipeline included Random Forest, Gradient Boosting, stacking, isotonic probability calibration, threshold-sensitivity analysis, and SHAP-based interpretation. The analysis was conducted in a notebook-based environment using a fixed random seed of 42 to improve computational reproducibility. The publicly available Student Depression Dataset obtained from Kaggle [33] was used throughout the experiments, and the implementation followed the preprocessing, validation, calibration, and interpretability procedures described in this section.

Primary model evaluation was performed using stratified five-fold outer cross-validation with nested hyperparameter optimization. Within each outer fold, the Random Forest and Gradient Boosting hyperparameters were selected exclusively from the corresponding training partition using an inner stratified cross-validation loop. The stacking ensemble was then constructed within the outer training fold using scikit-learn's internal cross-validated meta-feature generation procedure with cv = 5. Consequently, the Logistic Regression meta-learner was trained on cross-validated base-model predictions rather than on predictions obtained from the outer validation data.

Probability calibration was also performed without using the outer validation samples during model fitting. Isotonic calibration was fitted within the corresponding outer-fold training data, after which calibrated predictions were generated for the unseen outer validation fold. Predictions from all outer validation folds were subsequently combined to form a pooled out-of-fold prediction set. ROC curves, precision–recall curves, calibration plots, confusion matrices, threshold-sensitivity analyses, and pooled performance measures were calculated using these predictions. Thus, each observation contributed to the reported evaluation only when it belonged to an unseen outer validation partition.

Model performance was assessed using ROC–AUC, PR–AUC, recall, precision, F1-score, Brier score, and ECE. ROC–AUC was used to measure overall discrimination, while PR–AUC was included because it provides a screening-relevant view of precision and recall. Recall was treated as an important operational measure because false-negative predictions may prevent higher-risk students from being prioritized for further assessment. Precision and F1-score were reported to show the trade-off between identifying more positive cases and limiting unnecessary referrals. Accuracy was not emphasized because it can conceal class-specific performance and does not directly reflect probability reliability.

Threshold-sensitivity analysis was performed by applying different classification thresholds to the calibrated probabilities. This analysis was used to examine how lower or higher thresholds altered precision, recall, and F1-score under different screening policies. Lower thresholds represent higher-coverage screening strategies, whereas higher thresholds may be considered when institutional follow-up resources are limited.

SHAP was used to examine global and local feature contributions. Post-hoc visual explanations were generated using a Gradient Boosting model retrained on the full dataset after completion of the cross-validated performance evaluation. A randomly sampled background dataset of 100 observations and an evaluation subset of up to 500 observations were used with interventional feature perturbation. To assess explanation stability, SHAP feature rankings were also calculated separately within each outer validation fold, and their consistency was evaluated using feature frequency and pairwise Spearman rank correlation. The full-data SHAP visualizations were used only for interpretation and were not used to estimate predictive performance.

The overall experimental procedure consisted of the following steps:

1. Load the student depression dataset
2. Filter records to technical education students
3. Remove leakage-prone and non-deployable variables
4. Encode categorical and ordinal variables
5. Construct feature matrix and target labels
6. Perform outer stratified five-fold cross-validation
7. For each outer training fold:
   a. Perform inner cross-validation hyperparameter tuning
   b. Train Random Forest and Gradient Boosting models
   c. rain the stacking ensemble using scikit-learn's internal cross-validated meta-feature generation mechanism
   d. Apply isotonic probability calibration within the corresponding training partition
   e. Generate calibrated validation predictions for pooled out-of-fold evaluation
8. Generate pooled out-of-fold predictions
9. Compute discrimination and calibration metrics
10. Perform threshold sensitivity analysis

11. Conduct leakage-ablation and reduced-model analysis
12. Evaluate SHAP interpretability and stability across folds

3. Results

This section presents the empirical evaluation of the proposed leakage-aware and calibration-aware framework for early depression-risk screening. Model performance was assessed using stratified cross-validation, probability-calibration analysis, leakage-ablation experiments, threshold-sensitivity evaluation, subgroup robustness assessment, and SHAP-based interpretability analysis.

3.1 Cross-Validated Model Performance

Stratified five-fold cross-validation was used to minimize the variance induced by the single train–test split, leading to a reliable estimation of the model performance in the presence of a moderate class imbalance. The cross-validated discrimination, calibration, and screening-oriented metrics are summarized in Table 6.

As shown in Table 6, discrimination performance remained stable and highly comparable across all evaluated models when leakage-aware feature constraints were applied, with ROC–AUC values ranging from 0.8680 to 0.8718 and PR–AUC values consistently close to 0.889. The best results in terms of Recall were obtained by Gradient Boosting and the stacking ensemble, which are suitable for sensitivity oriented identification of higher-risk students.

The calibrated stacking ensemble achieved the strongest balance between discrimination, recall, and calibration reliability among the evaluated approaches, with ROC–AUC = 0.8718 ± 0.0081, recall = 0.8408 ± 0.0100, Brier score = 0.1434 ± 0.0049, and ECE = 0.0207 ± 0.0085. However, the performance gains over Gradient Boosting remained relatively modest, suggesting that much of the predictive signal was already captured by the strong single learners. Consequently, the ensemble design contributed more toward stable recall-oriented behavior and calibration reliability than toward substantial discrimination improvement.

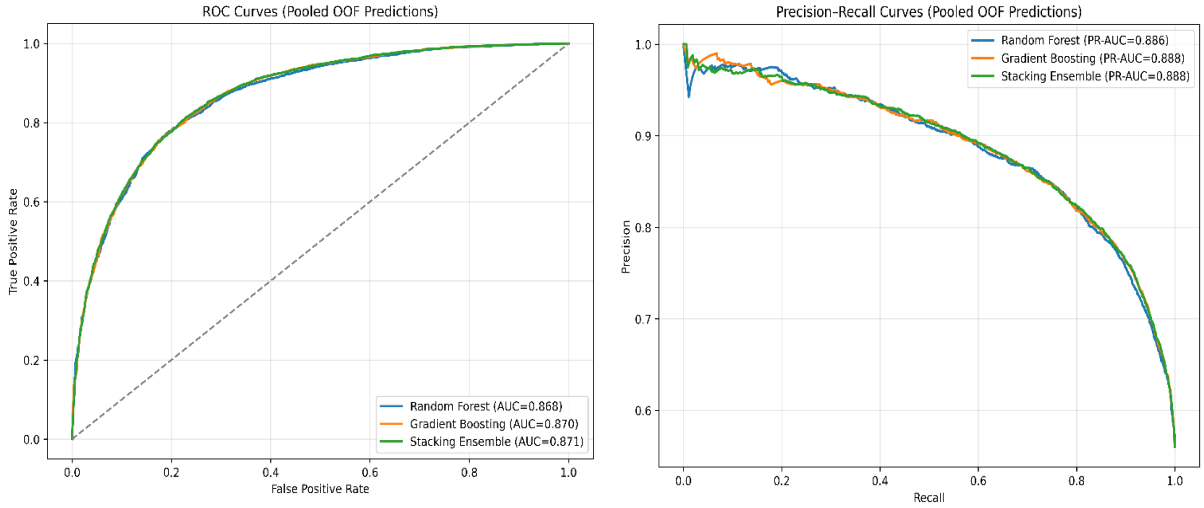
As summarized in Table 7, isotonic calibration improved probability reliability across all three evaluated models, as reflected by reductions in both Brier score and ECE. The calibration changes were greatest for Random Forest, with ECE decreasing from 0.0619 to 0.0241 after calibration. For Gradient Boosting and the stacking ensemble, calibration also improved, although the magnitude of improvement was smaller. The post-calibration ECE of the stacking ensemble was the lowest among the evaluated models.

Table 6. Cross-validated model comparison using discrimination, calibration, and screening-oriented metrics.

Model	ROC–AUC	PR–AUC	Recall	Precision	F1-score	Brier	ECE
Random Forest	0.8680 ± 0.0101	0.8868 ± 0.0097	0.8271 ± 0.0109	0.8013 ± 0.0079	0.8140 ± 0.0089	0.1456 ± 0.0058	0.0241 ± 0.0046
Gradient Boosting	0.8710 ± 0.0079	0.8883 ± 0.0072	0.8374 ± 0.0123	0.8007 ± 0.0075	0.8186 ± 0.0062	0.1439 ± 0.0047	0.0225 ± 0.0057
Stacking Ensemble	0.8718 ± 0.0081	0.8888 ± 0.0078	0.8408 ± 0.0100	0.8022 ± 0.0072	0.8210 ± 0.0051	0.1434 ± 0.0049	0.0207 ± 0.0085

Table 7. Pre- versus post-calibration reliability analysis

Model	Brier (Pre)	Brier (Post)	ECE (Pre)	ECE (Post)
Random Forest	0.1516 ± 0.0074	0.1456 ± 0.0058	0.0619 ± 0.0252	0.0241 ± 0.0046
Gradient Boosting	0.1445 ± 0.0049	0.1439 ± 0.0047	0.0289 ± 0.0087	0.0225 ± 0.0057
Stacking Ensemble	0.1443 ± 0.0051	0.1434 ± 0.0049	0.0368 ± 0.0059	0.0207 ± 0.0085



(a) ROC curve **(b) Precision–Recall**

Figure 2. ROC and Precision–Recall analysis of the evaluated models using pooled out-of-fold predictions (**a-b**).

A paired Wilcoxon signed-rank test across validation folds was used to determine whether the performance differences between Gradient Boosting and the stacking ensemble were statistically significant. The resulting p-values were p = 0.1250 for ROC–AUC, p = 1.0000 for PR–AUC, p = 0.0625 for calibrated Brier score, and p = 0.3125 for calibrated ECE. The results here suggest that there was no statistically significant improvement in discrimination from the ensemble over calibrated Gradient Boosting. Thus, the value of stacking in this context should not be interpreted as predictive superiority but rather as a potential improvement in robustness and probability reliability.

All visual analyses, such as ROC curves, precision–recall curves, calibration plots, confusion matrices, and threshold-sensitivity analyses were performed with pooled out-of-fold predictions, which were aggregated across validation folds.

Figure 2 presents the discrimination behaviour of Random Forest, Gradient Boosting, and the stacking ensemble using pooled out-of-fold predictions. The closely aligned ROC and precision–recall curves indicate comparable performance across the three models in identifying the positive depression-risk class under the observed class distribution.

3.2 Leakage Sensitivity Analysis

The effect of target leakage on predictive performance was examined by performing an ablation study that involved assessing performance of the model before and after the suicidal-thoughts variable was removed.

Table 8. Leakage-feature ablation analysis.

Configuration	ROC–AUC
GB (With Leakage, Uncalibrated)	0.9224
GB (No Leakage, Uncalibrated)	0.8701
Stacking (No Leakage, Uncalibrated)	0.8710
Stacking (No Leakage, Calibrated)	0.8713

The differing values of fold-averaged ROC–AUC reported in Table 6 and pooled out-of-fold ROC–AUC reported in Table 8 arise because the former are averaged across validation folds, whereas the latter are computed directly from pooled out-of-fold predictions across all samples.

ROC–AUC was artificially inflated from 0.8701 to 0.9224 by the leakage-prone feature, further demonstrating how symptom-level or post-outcome variables can substantially inflate predictive performance. This result emphasizes the need for leakage-aware feature selection in realistic early-screening systems and suggests that very high AUC values reported in some previous studies may partly reflect data leakage.

The ablation analysis showed that leakage-prone variables had a substantially larger effect on apparent performance than ensemble learning or calibration alone. Compared to the model without the suicidal-thoughts feature, the ROC–AUC decreased by around 0.05, while stacking and calibration on their own resulted in marginal discrimination gains. This finding suggests that temporally consistent feature design contributes more substantially to realistic screening performance than increasing architectural complexity.

3.3 Probability Calibration and Threshold Sensitivity Analysis

In screening-oriented systems, reliable probability estimation is critical as decisions to intervene are frequently made based on the probability of risk in addition to binary classifications. Thus, isotonic regression was used to improve the calibration of probability estimates produced by the stacking ensemble.

All threshold-specific analyses in this section were based on the probabilities from all outer validation folds that were pooled together. Thus, the reported precision–recall trade-offs are for fully unseen-sample evaluation and not a representative split.

Isotonic calibration, as illustrated in Figure 3(a), improved agreement between predicted probabilities and observed outcome frequencies, although moderate mismatches remained at intermediate confidence levels. Figure 3(b) shows the precision/recall tradeoff for various decision thresholds. Lowering the threshold increases sensitivity (recall) for identifying higher-risk students but also increases false-positive predictions, whereas higher thresholds improve precision at the cost of reduced recall. This behaviour will ensure that the system can be deployed flexibly in different institutional resource constraints. The threshold-specific precision, recall, and F1-score values are summarized in Table 9.

Table 9. Threshold-specific screening performance of the calibrated stacking ensemble

Threshold	Precision	Recall	F1-score
0.4	0.7716	0.9074	0.8340
0.5	0.8022	0.8408	0.8210
0.6	0.8550	0.7817	0.8167

The lower threshold (0.4) had a higher recall, and might therefore be preferable in high coverage screening settings where false negative minimization is the priority. On the other hand, setting the thresholds higher lowered the number of false-positive results, and might be more suitable in institutions with limited counseling or intervention capabilities.

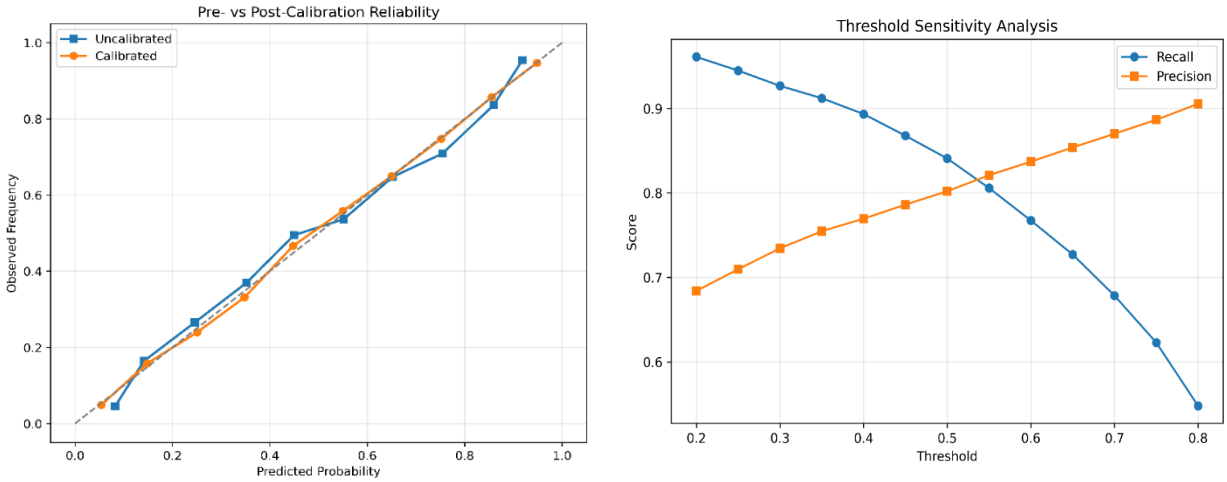
3.4 Reduced Screening Model and Subgroup Robustness

A less complicated model was tested that included only six predictors that were available at an early stage to assess feasibility of deployment. The performance of the reduced six-predictor model is reported in Table 10.

Despite using substantially fewer predictors, the reduced model maintained comparable discrimination and recall.

Table 10. Performance of the reduced-feature screening model.

Model	ROC–AUC	PR–AUC	Recall	Brier	ECE
Reduced screening model	0.8616	0.8761	0.8386	0.1484	0.0091



(a) Probability calibration **(b) Threshold sensitivity**

Figure 3. Calibration and threshold-sensitivity analysis of the stacking ensemble using pooled out-of-fold predictions (a-b).

Additionally, the reduced-feature configuration displayed lower observed ECE values in this evaluation, while the reduced model was not calibrated specifically for the evaluation. The results indicate that lightweight screening systems might be more operationally feasible to provide risk estimation and lessen dependence on elaborate psychosocial profiling.

A subgroup robustness analysis was also conducted across all eligible technical degree programs to assess the potential variability in predictive behavior across educational contexts. To simplify presentation, only the first three degree-specific subgroups eligible for evaluation within the implementation pipeline (B.Tech, BCA, and BE) were included, each satisfying minimum subgroup-size and label-variability requirements. Degree-wise ROC–AUC values are summarized in Table 11.

Table 11. Degree-wise subgroup robustness analysis

Degree Program	ROC–AUC
B.Tech	0.8790
BCA	0.8633
B.E	0.8291

The differences between degree programs indicate that there may be differences in mental health risk structures between technical disciplines. This finding indicates the need for more subgroup-specific assessment and suggests that general screening models may not fully reflect domain-specific stress patterns.

3.4.1 Class-Weight Robustness Analysis

To determine whether Gradient Boosting was sensitive to imbalance-aware weighting, the model was tested both with and without balanced sample weighting. The weighted and unweighted Gradient Boosting results are compared in Table 12.

Table 12. Class-weight robustness analysis

Configuration	ROC–AUC	PR–AUC	Recall	Brier
GB (Balanced Sample Weights)	0.8712	0.8883	0.7992	.1450
GB (Unweighted)	0.8714	0.8887	0.8388	0.1433

Results were very stable with or without class-weight adjustment, indicating that the proposed framework was not very sensitive to the assumptions of imbalance-handling. The analysis emphasized PR–AUC and recall because of their operational relevance to screening, rather than because of severe class imbalance, as the observed imbalance ratio was relatively modest (1.27).

3.5 Confusion Matrix and Interpretability Analysis

After cross-validated evaluation was completed, a Gradient Boosting model with the same hyperparameters was retrained on the full dataset for SHAP analysis. A randomly sampled background data set of 100 observations was used with the interventional feature perturbation. The encoded feature representation used during the model training was used to compute SHAP values on an evaluation subset of up to 500 samples that was randomly selected.

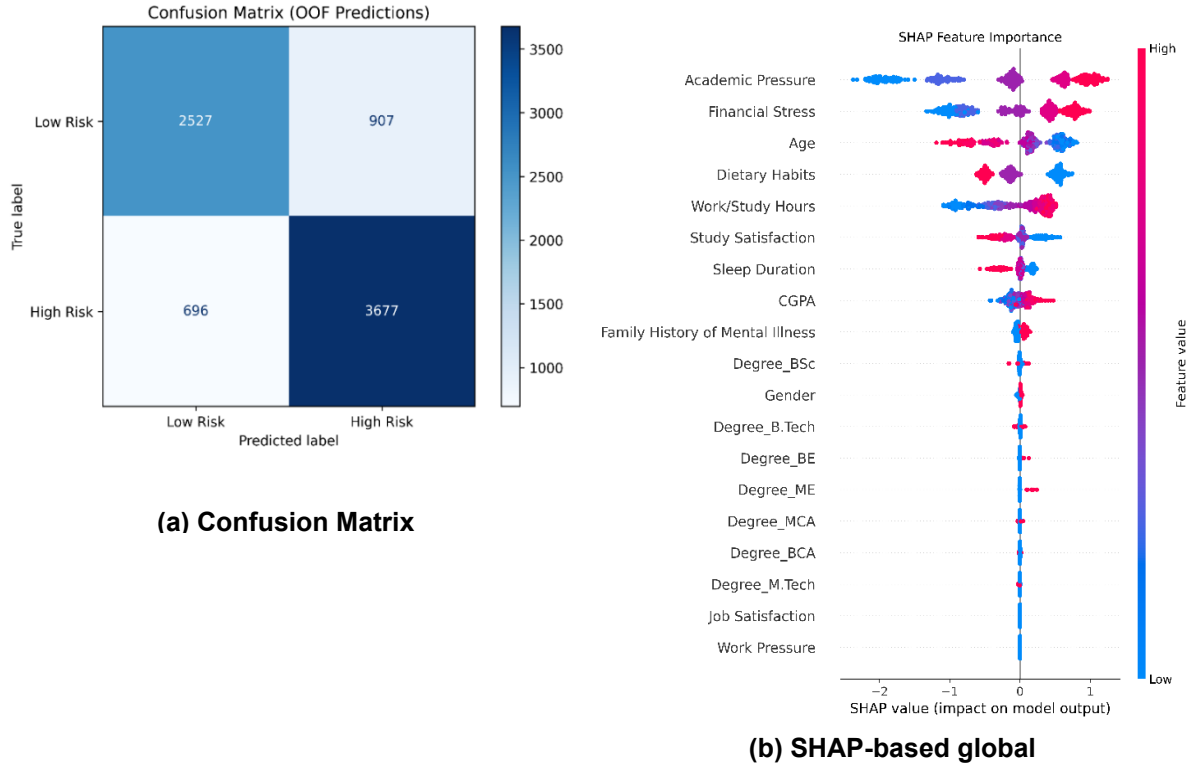
For robust evaluation of the interpretability, SHAP feature rankings were also calculated individually for each outer validation fold with a Gradient Boosting model for that fold. Stability was measured with feature-frequency analysis and with pairwise Spearman rank correlation.

Figure 4(a) shows that the confusion matrix yields good sensitivity in identifying higher risk students while keeping the precision at the default threshold level very competitive. The rest of the false negatives are due to the challenges of early-stage depression detection based on solely non-clinical and temporally consistent predictors.

The SHAP summary plot in Figure 4(b) shows that Academic Pressure and Financial Stress were the most influential contributors to predicted depression risk, followed by Work/Study Hours, Age, Dietary Habits, Sleep Duration, and CGPA. The results suggest that structural academic and socioeconomic pressures may have greater predictive influence than downstream lifestyle-related variables in technical education populations.

As summarized in Table 13, nine of the ten leading features appeared in all five validation folds, while Gender appeared in four folds. The mean pairwise Spearman rank correlation across folds was 0.9688, indicating highly consistent feature-importance rankings across validation partitions.

The local SHAP explanation shown in Fig. 5 illustrates the contribution of each predictor to the overall risk assessment of a typical high-risk student. Such local explanations may assist human-centred review and institutional decision support, but they should not be used for fully automated intervention systems.



(a) Confusion Matrix

(b) SHAP-based global

Figure 4. (a) Confusion matrix obtained from pooled out-of-fold predictions and (b) global SHAP feature importance obtained from the full-data Gradient Boosting interpretation model.

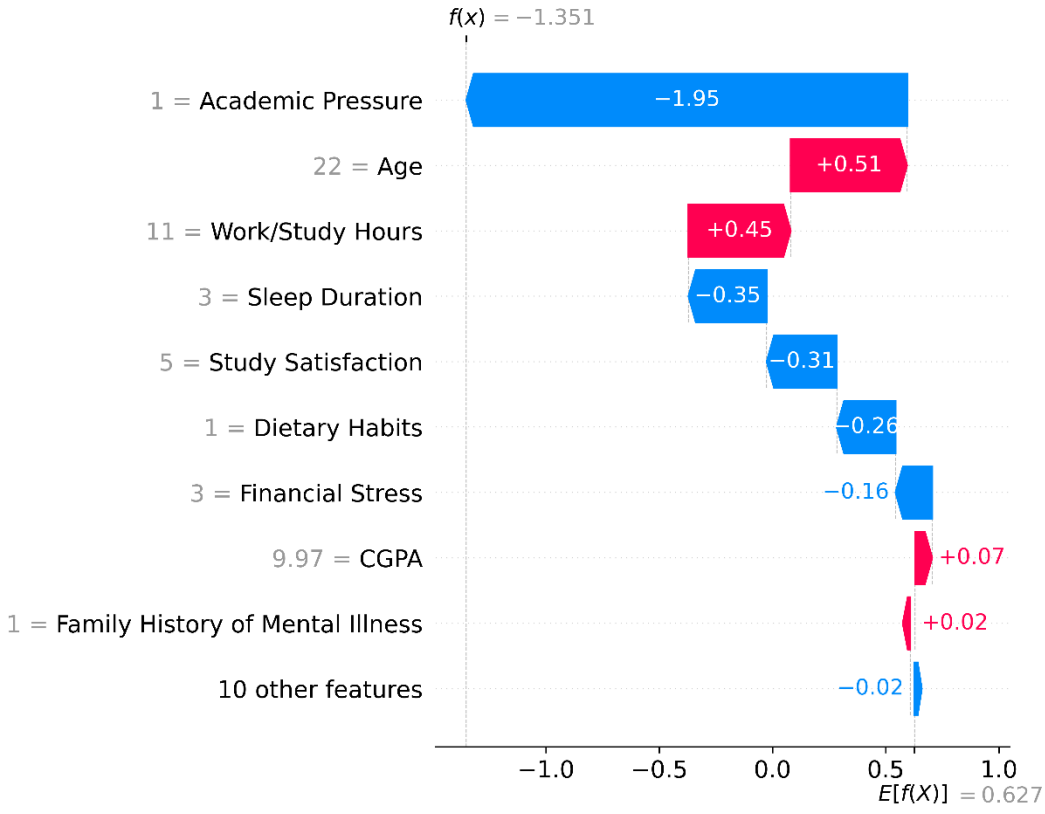


Figure 5. Instance-level SHAP explanation illustrating local feature contributions for a representative high-risk student.

Table 13. SHAP stability analysis across cross-validation folds.

Feature	Fold Frequency
Academic Pressure	5
Financial Stress	5
Work/Study Hours	5
Age	5
Dietary Habits	5
Study Satisfaction	5
Sleep Duration	5
CGPA	5
Family History of Mental Illness	5
Gender	4

4. Discussion

The results demonstrate that early depression-risk screening among technical education students can be achieved using leakage-aware and non-clinical predictors without substantially compromising discrimination performance, calibration reliability, or cross-validated robustness. The predictive performance observed in the present study is broadly consistent with recent machine learning-based student mental health studies reported by Bhattacharjee *et al.* [17] and Xu and Zhang [19], while the present framework additionally incorporates leakage-aware feature design, calibration assessment, and deployment-oriented validation. Unlike several earlier studies that primarily emphasized discrimination performance alone, the present study evaluates probability reliability and threshold-sensitive operational behavior alongside classification accuracy. Furthermore, the relatively low variation observed across validation folds suggests that the reported performance is unlikely to be attributable to favorable data partitioning or overfitting effects.

The stacking ensemble also resulted in slightly better recall, although statistical testing showed that the improvement over Gradient Boosting was not statistically significant. This observation is broadly consistent with prior healthcare-oriented machine learning studies reporting that ensemble-based approaches can improve prediction robustness and model stability in medical prediction tasks [7, 26]. Therefore, in the present framework, the principal advantage of stacking lies in its contribution to calibration consistency and screening robustness rather than dramatic improvements in ROC–AUC performance.

One of the most important findings of the study is the leakage-ablation analysis. The inclusion of the suicidal-thoughts variable substantially increased ROC–AUC performance, demonstrating how symptom-level or post-outcome predictors can artificially inflate predictive estimates. This finding is consistent with broader methodological concerns regarding leakage-prone predictors and unrealistic evaluation practices in healthcare-oriented machine learning systems [9, 10]. Consequently, the present results emphasize that temporally consistent feature design and leakage-aware evaluation represent important considerations for developing clinically and operationally realistic early-screening frameworks.

Interpretability analysis further suggested that depression-risk patterns within technical education populations may be associated with domain-specific academic and socioeconomic stressors. Academic Pressure and Financial Stress consistently appeared among the most influential predictors, which aligns with prior psychological and student mental-health studies reporting associations between academic burden, financial difficulties, and deteriorating student well-being [4, 5, 15, 16]. Sleep Duration also contributed consistently to model predictions, although its global importance was lower than that of Academic Pressure and Financial Stress. These findings suggest that institutional and socioeconomic factors may be relevant contributors to elevated depression-risk predictions within technical education cohorts.

The framework also supports configurable deployment through threshold-sensitive operation. Institutions with a goal of achieving high coverage for screening may decide to use a lower threshold, knowing that this will lead to more false-positive referrals, while institutions with limited counselor capacity may choose a higher threshold to minimize false-positive referrals. The proposed system should therefore be regarded not as an automated diagnostic system but as a flexible risk-prioritization tool.

However, there are a number of caveats to note. The study is based on self-reported survey data, and

response bias and measurement error may be a concern. Second, it is cross-sectional and thus unable to show temporal development of depressive symptoms. Third, although explicit leakage-prone variables were removed, some retained academic variables may still reflect latent psychological distress.

The framework was also evaluated using a single public dataset, and additional validation across diverse institutional, demographic, and cultural contexts is necessary to improve generalizability. Real-world deployment performance may additionally be affected by covariate shift, label shift, and measurement inconsistencies commonly observed in healthcare-oriented machine learning systems [7, 9]. Future research should therefore prioritize external multi-institutional validation, longitudinal assessment, and calibration-transfer evaluation under domain-shift conditions to further establish deployment reliability.

Despite these limitations, the proposed system offers a transparent, interpretable, and scientifically sound method for early depression risk screening. Instead of focusing on maximizing the classification accuracy, the study focuses on leakage-aware modeling, calibration reliability evaluation, robustness evaluation, and deployment-oriented assessment for responsible student mental health analytics.

5. Conclusion

This study proposes a leakage-aware machine-learning framework for early depression-risk screening among technical education students using institutionally accessible predictors and calibrated probability estimates. Stable cross-validated performance was observed when the method was evaluated on a filtered sample of 7,807 students, with ROC–AUC close to 0.87 and PR–AUC close to 0.89 in a pooled out-of-fold evaluation. While the calibrated stacking ensemble obtained slightly better overall performance, statistical testing revealed no statistically significant difference between the calibrated stacking and the calibrated Gradient Boosting, suggesting that, for this dataset, more complex architectures did not significantly improve predictive performance. Leakage-ablation analysis also revealed the critical need for realistic feature selection because the post-outcome suicidal-thoughts variable significantly boosted the apparent performance of the algorithm. Calibration and threshold-sensitivity analyses were used to support deployment-oriented risk prioritization, while SHAP analysis identified academic pressure and financial stress as influential factors across validation folds. The proposed framework provides an interpretable, deployment-oriented, and methodologically transparent basis for early depression risk screening in an institutional context based on non-clinical student data.

References

[1] World Health Organization. (2022). World Mental Healthreport: Transforming Mental Health for all. Executive Summary. World Health Organization.

[2] P. Akhtar, L. Ma, A. Waqas, S. Naveed, Y. Li, A. Rahman, Y. Wang, Prevalence of depression among university students in low and middle income countries (LMICs): a systematic review and meta-analysis. Journal of Affective Disorders, 274, (2020) 911–919. https://doi.org/10.1016/j.jad.2020.03.183

[3] R.P. Auerbach, P. Mortier, R. Bruffaerts, J. Alonso, C. Benjet, P. Cuijpers, D.D. Ebert, J.G. Green, I. Hasking, E. Murray, R.C. Kessler, Mental Disorders Among College Students in the World Health Organization World Mental Health Surveys. Psychological Medicine, 46(14), (2016) 2955–2970. https://doi.org/10.1017/S0033291716001665

[4] T. Richardson, P. Elliott, R. Roberts, The Relationship Between Personal Unsecured Debt and Mental and Physical Health: A Systematic Review and Meta-Analysis. Clinical Psychology Review, 33(8), (2013) 1148–1162. https://doi.org/10.1016/j.cpr.2013.08.009

[5] T. Richardson, P. Elliott, R. Roberts, M. Jansen, A Longitudinal Study of Financial Difficulties and Mental Health in a National Sample of British Undergraduate Students. Community Mental Health Journal, 53(3), (2017) 344–352. https://doi.org/10.1007/s10597-016-0052-0

[6] A.B.R. Shatte, D.M. Hutchinson, S.J. Teague, Machine Learning in Mental Health: A Scoping Review of Methods and Applications. Psychological Medicine, 49(9), (2019) 1426–1448. https://doi.org/10.1017/S0033291719000151

[7] A. Rajkomar, J. Dean, I. Kohane, Machine Learning in Medicine. New England Journal of Medicine, 380(14), (2019) 1347–1358. https://doi.org/10.1056/NEJMra1814259

[8] C. Yu, Y. Zhang, H. Liu, X. Li, Machine Learning Models for Predicting the Risk of Depressive Symptoms in Chinese College Students. Frontiers in Psychiatry, 16, (2025) 1648585. https://doi.org/10.3389/fpsyt.2025.1648585

[9] I.Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, M. Ghassemi, Ethical Machine Learning in Healthcare. Annual Review of Biomedical Data Science, 4, (2021) 123–144. https://doi.org/10.1146/annurev-biodatasci-092820-114757

[10] M.A. Hernán, J.M. Robins, (2020). Causal Inference: What If, Chapman and Hall/CRC, Boca Raton, FL, USA.

[11] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On Calibration of Modern Neural Networks.

Proceedings of the 34th International Conference on Machine Learning (ICML), 70, (2017) 1321–1330.

[12] M. Kull, M. Perello-Nieto, M. Kängsepp, T. Silva Filho, H. Song, P. Flach, Beyond Temperature Scaling: Obtaining Well-Calibrated Multiclass Probabilities with Dirichlet Calibration. Advances in Neural Information Processing Systems, 32, (2019) 12295–12305.

[13] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. Information Fusion, 58, (2020) 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[14] C. Molnar, (2020). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Germany: Leanpub.

[15] G. Barbayannis, M. Bandari, X. Zheng, H. Baquerizo, K.W. Pecor, X. Ming, Academic Stress and Mental Well-Being in College Students: Correlations, Affected Groups, and COVID-19. Frontiers in Psychology, 13, (2022) 886344. https://doi.org/10.3389/fpsyg.2022.886344

[16] D.C. Jessop, M. Reid, L. Solomon, Financial Concern Predicts Deteriorations in Mental and Physical Health Among University Students. Psychology and Health, 35(2), (2020) 196–209. https://doi.org/10.1080/08870446.2019.1626393

[17] S. Bhattacharjee, M.F, Hossain, S.A. Akhy, M.M.H. Shimul, M.K. Hossain, Predicting Depression Among Public University Students in Bangladesh using Machine Learning Models, Discover Mental Health, 5, (2025) 191. https://doi.org/10.1007/s44192-025-00332-0

[18] R.P. Sonjaya, A.R. Gintara, L.S. Riza, M. Nursalman, E. Nugraha, D. Wahyudin, Predicting Student Depression using the Naive Bayes Model on the Student Depression Dataset from Kaggle. JENTIK: Jurnal Pendidikan Teknologi Informasi dan Komunikasi, 4(1), (2025) 9–20. https://doi.org/10.58723/jentik.v4i1.448

[19] X. Xu, T. Zhang, Intelligent Classification and Prediction of Students' Mental Health in Online Learning Environments using Boosting Algorithm and LIWC Features. Scientific Reports, 15, (2025) 22028. https://doi.org/10.1038/s41598-025-04681-2

[20] M. Gil, S.-S. Kim, E. J. Min, Machine Learning Models for Predicting Risk of Depression in Korean College Students: Identifying Family and Individual Factors. Frontiers in Public Health, 10, (2022) 1023010. https://doi.org/10.3389/fpubh.2022.1023010

[21] N. Meda, S. Pardini, P. Rigobello, F. Visioli, C. Novara, Frequency and Machine Learning Predictors of Severe Depressive Symptoms and Suicidal Ideation Among University Students. Epidemiology and Psychiatric Sciences, 32, (2023) e42. https://doi.org/10.1017/s2045796023000550

[22] A. Daza Vergaray, J.C. Herrera Miranda, J. Bobadilla Cornelio, A.R. López Carranza, C.F. Ponce Sánchez, Predicting the Depression in University Students using Stacking Ensemble Techniques Over Oversampling Method. Informatics in Medicine Unlocked, 41, (2023) 101295. https://doi.org/10.1016/j.imu.2023.101295

[23] A.H. Chowdhury, D. Rad, M.S. Rahman, Predicting Anxiety, Depression, and Insomnia Among Bangladeshi University Students using Tree-Based Machine Learning Models. Health Science Reports, 7(4), (2024) e2037. https://doi.org/10.1002/hsr2.2037

[24] L. Zhang, S. Zhao, Z. Yang, H. Zheng, M. Lei, An Artificial Intelligence Tool to Assess the Risk of Severe Mental Distress Among College Students in Terms of Demographics, Eating Habits, Lifestyles, and Sport Habits: An Externally Validated Study using Machine Learning. BMC Psychiatry, 24(1), (2024) 581. https://doi.org/10.1186/s12888-024-06017-2

[25] M.E. Hasan, M. Arif, S.M.R. Hasan, M. Muwanguzi, J. Abaatyo, M.M. Kaggwa, M.M. Almerab, P.A. Atroszko, M. Muhit, F. Al-Mamun, M.A. Mamun, Prevalence, Associated Factors, and Machine Learning-Based Prediction of Depression, Anxiety, and Stress Among University Students: A Cross-Sectional Study from Bangladesh. Journal of Health, Population and Nutrition, 44(1), (2025) 361. https://doi.org/10.1186/s41043-025-01095-8

[26] S. Uddin, A. Khan, M.E. Hossain, M.A. Moni, Comparing Different Supervised Machine Learning Algorithms for Disease Prediction. BMC Medical Informatics and Decision Making, 19(1), (2019) 281. https://doi.org/10.1186/s12911-019-1004-8

[27] H. He, Y. Ma, (2013). Imbalanced Learning: Foundations, Algorithms, and Applications, Wiley-IEEE Press. https://doi.org/10.1002/9781118646106

[28] T. Saito, M. Rehmsmeier, The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE, 10(3), (2015) e0118432. https://doi.org/10.1371/journal.pone.0118432

[29] C. Mimikou, C. Kokkotis, D. Tsiptsios, K. Tsamakis, S. Savvidou, L. Modig, F. Christidi, A. Kaltsatou, T. Doskas, C. Mueller, A. Serdari, Explainable Machine Learning in the Prediction of Depression. Diagnostics, 15(11), (2025) 1412.

https://doi.org/10.3390/diagnostics15111412

[30] E.J. Topol, (2019). The Topol Review: Preparing the Healthcare Workforce to Deliver the Digital Future, Health Education England, United Kingdom.

[31] U. Madububambachu, A. Ukpebor, U. Ihezue, Machine Learning Techniques to Predict Mental Health Diagnoses: A Systematic Literature Review. Clinical Practice and Epidemiology in Mental Health, 20, (2024) e17450179315688. https://doi.org/10.2174/0117450179315688240607052117

[32] M. Drira, S. Ben Hassine, M. Zhang, S. Smith, Machine Learning Methods in Student Mental Health Research: An Ethics-Centered Systematic Literature Review. Applied Sciences, 14(24), (2024) 11738. https://doi.org/10.3390/app142411738

[33] Shodolamu Opeyemi. (2024). Student Depression Dataset. Kaggle.Com. https://www.kaggle.com/datasets/hopesb/student-depression-dataset

Authors Contribution Statement

Mohitsinh Parmar: Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Investigation, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing. Sohil D. Pandya: Conceptualization, Methodology, Validation, Project Administration, Supervision, Writing – Review & Editing. Both authors have read and agreed to the published version of the manuscript.

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

Has this article screened for similarity?

Yes

About the License