Explainable Stacking Ensemble with Fairness-Aware Analysis for Student Depression Prediction

Nezar Abdilah Prakasa

Master of Artificial Intelligence, Universitas Gadjah Mada, Yogyakarta, Indonesia
nezarabdilahprakasa@mail.ugm.ac.id

Abstract -- Student depression has become a critical public health concern in higher education. Prior studies on the Kaggle Student Depression Dataset employed single-model classifiers without addressing class imbalance, interpretability, or subgroup fairness. This paper proposes an explainable stacking ensemble integrating Random Forest, Gradient Boosting, and XGBoost as base learners with Logistic Regression as meta-learner, augmented by SMOTE, SHAP, and fairness analysis across gender and financial stress subgroups. Experiments on 27,870 student records show the stacking ensemble achieves best Accuracy (0.8461) and F1-Score (0.8689). SHAP identifies suicidal ideation, academic pressure, and financial stress as dominant predictors. Fairness analysis uncovers a Recall gap of 0.272 between low- and high-financial-stress students, a novel finding with direct clinical implications.

Keywords: student depression, stacking ensemble, explainable AI, SHAP, fairness analysis, SMOTE, machine learning

I. INTRODUCTION

Depression among university students has emerged as a pressing global concern. Academic pressure, financial difficulties, and poor sleep habits are significant contributors to deteriorating mental health in higher education [1]. Early and accurate detection is essential to enable timely counseling intervention.

Machine learning (ML) approaches have shown considerable promise for predicting student depression from structured survey data. The Kaggle Student Depression Dataset [2] has become a standard benchmark comprising 27,901 records with features spanning academic, psychological, and lifestyle dimensions.

However, existing studies exhibit three critical limitations. First, prior work [3] employs single-model classifiers without addressing class imbalance, resulting in low Sensitivity (Recall = 0.685). Second, although [4] incorporates SHAP for feature importance, explanations are not translated into actionable insights. Third, no existing study evaluates fairness across demographic or socioeconomic subgroups, a critical gap for mental health prediction.

To address these gaps, this paper makes four contributions: **(1)** A stacking ensemble (RF+GB+XGB+LR meta-learner) achieving best Accuracy (0.8461) and F1-Score (0.8689). **(2)** SMOTE oversampling for class balancing. **(3)** SHAP-based explainability identifying suicidal ideation, academic pressure, and financial stress as dominant predictors. **(4)** The first fairness-aware evaluation on this dataset, revealing a Recall gap of 0.272 across financial stress subgroups.

II. RELATED WORK

Early ML work on this dataset [3] applied Naive Bayes and Decision Trees, achieving Specificity of 84.21% but Sensitivity of only 68.42%, indicating systematic failure to detect true depression cases. Class imbalance was not addressed.

More recent work [4] compared LR, RF, and XGBoost with SHAP, reporting best AUC of 0.913. While SHAP was employed, no fairness analysis was conducted and explanations remained at the population level without actionable output.

A systematic review [5] found XGBoost achieving best average performance (F1: 0.86, AUC: 0.84), with SHAP as the dominant interpretability method in 70% of studies. Performance differences across algorithms were not statistically significant ($p > 0.05$).

Stacking ensemble methods [6] leverage complementary inductive biases of diverse base learners through a meta-learner trained on out-of-fold predictions, consistently improving tabular classification. The I-HOPE framework [8] identified black-box dominance and lack of personalized insights as persistent gaps in student mental health ML.

III. DATASET AND METHODOLOGY

A. Dataset

The Student Depression Dataset [2] (Kaggle: hopesb/student-depression-dataset) comprises 27,901 records. After filtering to Profession='Student', 27,870 valid records remain with 17 features covering demographic, academic, psychological, and lifestyle dimensions. The binary target Depression is mildly imbalanced: 58.5% positive vs. 41.5% negative.

Three records with missing Financial Stress values were median-imputed. Profession and City were dropped. Categorical variables were Label-Encoded. An 80:20 stratified split (random_state=42) yielded 22,296 training and 5,574 test samples.

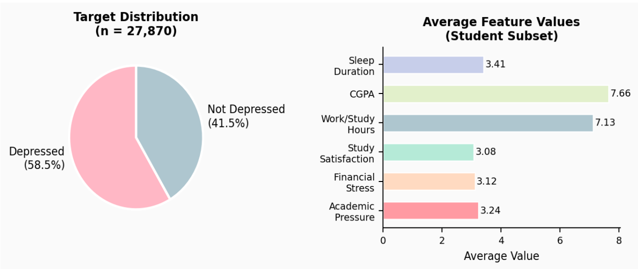

Fig. 1. Dataset overview: class distribution and average feature values.

B. Preprocessing and Class Balancing

SMOTE [9] was applied to the training set to balance classes (13,046 depressed vs. 9,250 non-depressed), producing 13,046 samples per class. Features were standardized using StandardScaler fitted on the resampled training set and applied identically to the test set to prevent data leakage.

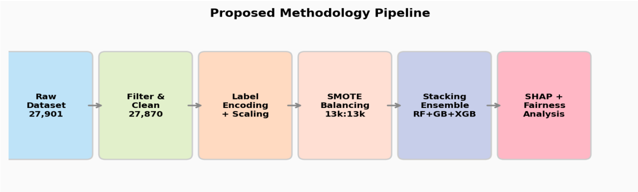

Fig. 2. Methodology pipeline from raw data to fairness evaluation.

C. Stacking Ensemble Architecture

Four baselines are evaluated: Logistic Regression (LR), Random Forest (RF, 100 trees), Gradient Boosting (GB, 100 estimators), and XGBoost (XGB, 100 estimators). The proposed Stacking Ensemble uses RF, GB, and XGB as heterogeneous base learners with 5-fold CV generating out-of-fold predictions, then trains LR as the meta-learner.

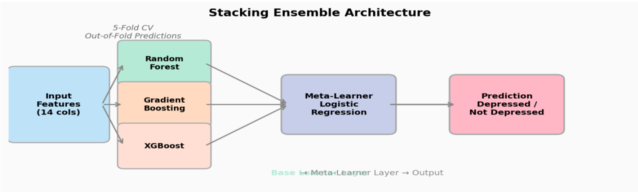

Fig. 3. Stacking ensemble architecture with 5-fold cross-validation.

D. Explainability and Fairness

SHAP TreeExplainer is applied to XGBoost across 5,574 test samples. Fairness analysis evaluates Accuracy, Recall, and F1-Score per Gender group and per Financial Stress level (1-5). Subgroup Recall is prioritized clinically since low Recall means depressed students are systematically missed.

IV. EXPERIMENTAL RESULTS

A. Model Performance Comparison

Table I presents all model metrics on the 5,574-sample test set. The Stacking Ensemble achieves the highest Accuracy (0.8461) and F1-Score (0.8689). Gradient Boosting attains best AUC-ROC (0.9207), while XGBoost achieves highest Recall (0.8786). Performance spread is narrow within 0.6%, consistent with [5].

Model	Acc	Prec	Rec	F1	AUC
Log. Regression	0.8437	0.8702	0.8614	0.8658	0.9202
Random Forest	0.8405	0.8581	0.8716	0.8648	0.9149
Grad. Boosting	0.8443	0.8643	0.8706	0.8674	0.9207
XGBoost	0.8423	0.8558	0.8786	0.8670	0.9127
Stacking (Ours)	**0.8461**	**0.8660**	**0.8719**	**0.8689**	**0.9202**

TABLE I. Model comparison. Bold row = proposed stacking ensemble.

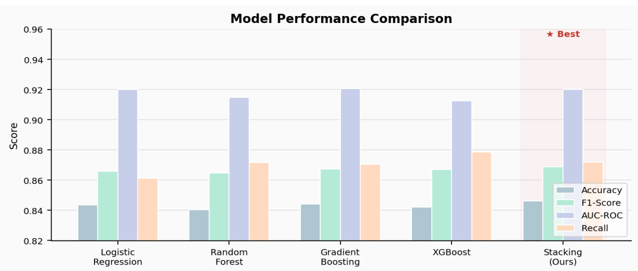

Fig. 4. Performance comparison across all models and four evaluation metrics.

The stacking ensemble combines RF's variance reduction, GB's sequential error correction, and XGB's regularized boosting, with LR as meta-learner learning optimal combination weights, yielding the most balanced multi-metric profile.

B. SHAP Explainability Analysis

Fig. 5 presents SHAP global feature importance from XGBoost. Suicidal Thoughts dominates (mean |SHAP|=1.42), followed by Academic Pressure (1.05) and Financial Stress (0.74). These three features account for the majority of predictive signal. Study Satisfaction acts as a protective factor with negative SHAP direction.

Directional analysis shows high Academic Pressure and Financial Stress consistently push predictions toward depression. Younger Age associates with higher depression risk. Work Pressure and Job Satisfaction show near-zero SHAP impact, consistent with the student-focused dataset context.

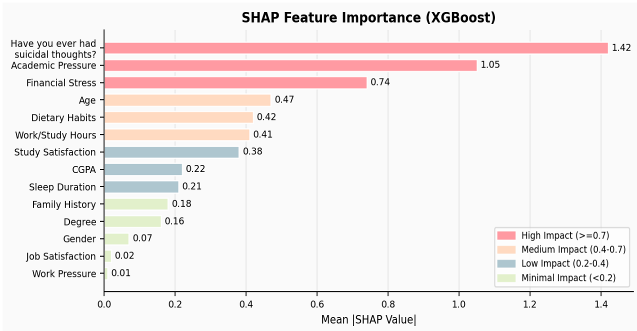

Fig. 5. SHAP feature importance (XGBoost). Darker bars indicate higher impact tiers.

C. Fairness Analysis

Table II presents subgroup performance. For Gender, the model is equitable: Female and Male both attain F1=0.867 with minimal Recall difference (0.882 vs. 0.876), indicating no gender-based disparate impact.

Subgroup	n	Acc	Recall	F1
Gender: Female	2,456	0.842	0.882	0.867
Gender: Male	3,118	0.843	0.876	0.867
Fin. Stress: Lv. 1	**994**	**0.827**	**0.685**	**0.716**
Fin. Stress: Lv. 2	1,014	0.821	0.783	0.792
Fin. Stress: Lv. 3	1,045	0.825	0.860	0.854
Fin. Stress: Lv. 4	1,148	0.855	0.916	0.897
Fin. Stress: Lv. 5	**1,373**	**0.872**	**0.957**	**0.922**

TABLE II. Fairness by subgroup. Dark=underdetection (Lv.1); light=best (Lv.5).

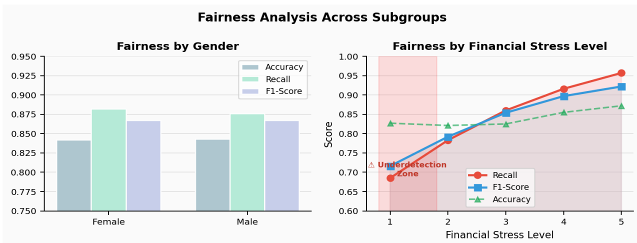

Fig. 6. Fairness: gender parity (left) and financial stress gradient (right).

Financial Stress reveals a critical gradient: Recall rises from 0.685 (Level 1) to 0.957 (Level 5), a gap of 0.272. Depressed students with low financial stress are systematically underdetected, likely driven by academic perfectionism or social isolation. No prior study has identified this disparity.

V. DISCUSSION

The narrow performance margins across models indicate the dataset is well-structured. The stacking ensemble's primary value lies in a more robust, balanced prediction profile rather than dramatic accuracy gains over individual models.

The SHAP finding that Suicidal Ideation is the strongest predictor raises ethical considerations. Automated screening systems relying on this feature require mandatory human-in-the-loop review with appropriate institutional safeguards to ensure student safety and privacy.

The Financial Stress fairness gap is the most actionable finding. Universities should recognize that low-financial-stress depressed students are more likely to be missed. Interventions targeting academic pressure and social well-being, beyond financial aid, are necessary to reach this underdetected population.

Limitations: (1) no longitudinal measurements; (2) geographic scope limited to Indian universities; (3) fairness covers only two attribute dimensions. Future work should validate in Southeast Asian contexts including Indonesian higher education.

VI. CONCLUSION

This paper presented an explainable stacking ensemble for student depression prediction using 27,870 Kaggle records, integrating SMOTE, stacking ensemble learning, SHAP explainability, and the first fairness-aware subgroup evaluation on this dataset.

Key contributions: (1) Stacking achieves best Accuracy (0.8461) and F1-Score (0.8689); (2) SHAP identifies suicidal ideation, academic pressure, and financial stress as dominant predictors; (3) Fairness analysis uncovers a Recall gap of 0.272 across financial stress levels, a novel clinically significant finding demonstrating systematic underdetection of low-stress depressed students. Future work will extend to longitudinal data and Southeast Asian university populations.

REFERENCES

[1] World Health Organization, "Mental health of adolescents," WHO Fact Sheets, 2021.

[2] hopesb, "Student Depression Dataset," Kaggle, 2024. Available: kaggle.com/datasets/hopesb/student-depression-dataset

[3] Anonymous, "Predicting Student Depression Using the Naive Bayes Model," ResearchGate, 2026.

[4] J. Peng et al., "Predicting Student Depression Using Machine Learning," AIDML 2025.

[5] S. Ahmad et al., "Systematic Review: Explainable ML for Clinical Depression Detection," PMC, 2025.

[6] D. H. Wolpert, "Stacked Generalization," Neural Networks, vol. 5, no. 2, pp. 241-259, 1992.

[7] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning. MIT Press, 2023.

[8] Y. Li et al., "Predicting College Student Mental Health with Interpretable ML," arXiv:2503.08002, 2025.

[9] N. V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," JAIR, vol. 16, pp. 321-357, 2002.