

Actionable and Trustworthy AI for Student Depression Prediction Using Multi-Level Explainability

Sanjit Kumar Dash¹, Deepak Kumar Pradhan², Mohammed Altaf Ahmed^{3*}, Qutubuddin Mohammed⁴

Odisha University of Technology and Research, Bhubaneswar, India^{1,2}

Department of Computer Engineering, College of Computer Engineering and Sciences,

Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia³

Department of Electrical Engineering, College of Engineering, Wadi Addawasir,

Prince Sattam bin Abdulaziz University, Wadi Addawasir, Saudi Arabia⁴

Abstract—It is critical to be able to identify and intervene with high-risk students as early as possible - this is most easily accomplished by developing scalable systems for providing such early warning signals. Although many researchers have developed highly accurate machine learning (ML) models to detect depression, poor clinical transparency makes it challenging to get adoption from both educators and physicians. This study presents a socio-technical framework addressing the ethical and clinical issues arising from using 'black-box' predictive models for high-stakes student mental health screening within computational psychiatry. By applying a two-level stacking ensemble that combines Random Forest (RF), Support Vector Machine (SVM) and Logistic Regression (LR) as base learners with a LR meta-learner for optimizing predictive performance. Class imbalance typical of mental health data is mitigated using Synthetic Minority Over-sampling Technique (SMOTE). Embedded within this socio-technical system, a dual-layered Explainable Artificial Intelligence (XAI) subsystem of SHAP and DiCE explains the model's inductive biases and proposes individually actionable recourse to high-risk students, ensuring ethical clinical translation. Empirically, the model exhibits 0.8421 accuracy and 0.9184 ROC-AUC. This study contributes to the deployment of interpretable ML within computational psychiatry and presents an extensible ethical framework that offers both accurate mental health risk identification and individualizable mental health intervention for university students.

Keywords—Explainable AI (XAI); student depression; stacking ensemble; data security; SMOTE; SHAP; LIME; Counterfactual Explanations (DiCE)

I. INTRODUCTION

Mental disorders have become a global crisis, and university students are now the most affected group. Student mental disorders, including depression, anxiety, and psychological distress, are being increasingly widely acknowledged and studied around the world, still, are still inflicting on the student's academic performance, psychological health and college retention of higher education [1]. Current screening systems face issues such as low acceptance rates because of stigmatization and lack of resources and student involvement; therefore, there is a worldwide demand for innovative and scalable data-driven predictive models

providing early, real-time detection of problems and intervention pathways [2]. Besides prediction, such prediction models must also provide actionable and interpretable insights to every student. Most current prediction models on student mental health lack a single complex algorithm like Random Forests to act as a "black-box" model [3]. In the area of student mental health, a black-box model will lead to distrust since no stakeholder would be confident about their predictions. Such issues, along with limited capabilities in addressing class imbalance in student mental health data and in generating diverse individualised counterfactual explanations that guide individual decisions, reveal the urgency of developing intelligent and interpretable prediction models for all students in higher education [4].

The presented work aims to alleviate the problem by presenting an ensemble model (using the Stacking classifier) based on an improved version of SMOTE: a single ensemble classifier is built with base classifiers such as Random Forest, Support Vector Machine, and Logistic Regression, and Logistic Regression is used as the meta-classifier. The use of ensembles generally is to potentially increase prediction accuracy on imbalanced data compared to individual classifiers. However, for interpretability purposes, Shapley Additive exPlanations (SHAP) is used for global and local feature explanations, and for counterfactual recourse options, Diverse Counterfactual Explanations (DiCE) are used.

II. RELATED WORK

During the past few years, the application of machine learning in predicting mental health disorders has significantly improved from applying a single classifier model to using ensembles of classifiers. Previous studies have applied single algorithms, including RF and SVM, which were more understandable and provided an interpretative benchmark for assessing prediction capability [5]. The problems in prediction capabilities and generalizability to different groups of students called for further investigation into ensemble models. From 2024-2026, greater focus was placed on applying architecture designs such as XGBoost and stacking ensembles [6,7]. This approach involved applying a number of base learners, and it took advantage of the complementary nature of base learners.

*Corresponding author.

In regard to accuracy in prediction, such ensembles often had an AUROC value of 0.91.

Data imbalance has been a challenge that was commonly faced in relation to mental health datasets, specifically in relation to detecting cases of severe depression, which were part of the minority class. An immediate effect of performing classification on imbalanced data would be a high error rate in predicting the minority class, which poses challenges for the validity of the classifier's predictions [8]. It has been conclusively proven through studies carried out over the years that the application of Synthetic Minority Over-sampling Technique (SMOTE) provides solutions to problems posed by data imbalance by providing synthetic data for the minority class. SMOTE was applied in this study to deal with data imbalance in the case of 'Severe' depression instances [9].

Over the last several years (2024–2026), applying a multi-level XAI framework to the prediction of students' mental health has been an outstanding research stream. Previous models were referred to as "black boxes" as they gained high predictive power that did not correspond to the stakeholders' trust or usage of the model [10]. Thus, techniques such as SHapley Additive exPlanations (SHAP) can provide feature attributions at a global and subgroup level and an overall picture of the main model drivers; other techniques like Diverse Counterfactual Explanations (DiCE) and Local Interpretable Model-agnostic Explanations (LIME) provide explanations for individual behavior, which assist in action-oriented advice regarding students' mental health and behavioral changes. While a considerable amount of research has been done on this topic, a primary concern raised in the 2026 study on the issue was that explainability models offer limited actionable individual recourse that prevented translating interpretability to concrete behavior modifications and policy intervention recommendations [11, 12]. In this study, a SMOTE-enhanced Stacking Ensemble is combined with multi-level XAI. We provide global and subgroup-level explanations via SHAP and produce diverse and achievable counterfactual explanations via DiCE as an example of individual recourse.

Currently, the majority of the predictive models aimed at assessing student mental health utilize complex algorithms that are considered "black box" systems. As such, these predictive systems produce an output (i.e., a prediction of risk of mental health disorder) but do not provide an intuitive way to interpret how those predictions were derived or developed. Thus, the lack of interpretability of the prediction models leads to a significant break in the trust between involved stakeholders (i.e., students, clinicians, and university administrators), as it is unclear to any involved stakeholder whether they can trust the prediction models as their foundation for making decisions about how to support at-risk students. This overall lack of transparency of the predictive models also limits their overall clinical value as well as ethical deployment in student support services [13,14].

Data available for predicting depression in students is also very imbalanced; thus, a disproportionate number of cases of depression are classified as being considered "high risk"; however, classifying very few students as being "at risk" in fact

leads to high error rates in predicting students who are at risk of depression, thereby making the validity/reliability of predictions based on classification systems fundamentally lousy. Therefore, standard classification systems do not effectively represent the true positive rate for students who are truly at risk for becoming depressed, and therefore the clinical usefulness of predictive models in helping to prevent or treat the potential onset of depression in students is severely limited [15].

While there are existing systems that identify students who are "high risk" for developing depression, these systems do not provide at-risk students with actionable recommendations tailored specifically to their own individual situations. There are no mechanisms in place to create a diversity of individualized counterfactual explanations (i.e., information that identifies what specific changes a student could make to improve their predictive risk) for any student identified as "high risk". Therefore, predictions are essentially purely diagnostic as to whether or not the student may become depressed, but not therapeutic in nature for the development of intervention designs focused on students [16,17].

III. PROPOSED METHODOLOGY

A multi-stage pipeline that consists of a machine learning framework capable of robust prediction modelling and interpretability is the proposed machine learning framework used, as shown in Fig. 1. The workflow gradually progresses from data preparation and class balancing to a hierarchical stacking ensemble, concluding with an AI (XAI) layer that provides conversion of the predictive output into an easy-to-understand format for interpreters.

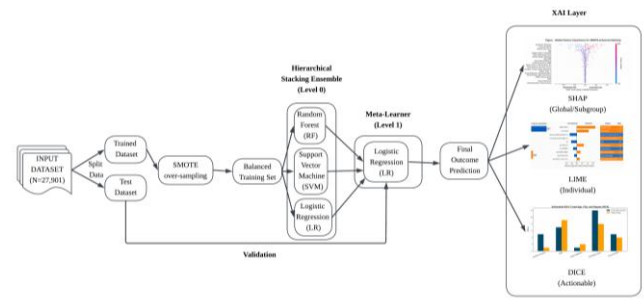


Fig. 1. Proposed methodology for multi-label XAI.

A. Dataset Description

The dataset that was used for this research effort consists of 27,901 total observations with 18 descriptive characteristics present in Table I. Due to the number of samples provided within this dataset, there is a solid basis for some type of statistical analysis.

The aim of this research is to predict outcomes from the Student Depression Dataset. This dataset includes the binary response variable 'Depressed' as one class and 'Non-depressed' as the second class. The classification algorithms will be trained to predict the response attribute based on the various available features, such as academic, demographic, and psychosocial features of the subjects. Note that the labels are taken as outcomes of the dataset and not as verified diagnostics or emerging depression status.

TABLE I. DATASET FEATURE DESCRIPTION

Feature	Type	Missing (%)
ID, Age, CGPA, Work/Study Hours	Numerical	0.00%
Gender, City, Profession, Sleep Duration, Dietary Habits, Degree	Categorical	0.00%
Academic Pressure, Work Pressure, Study Satisfaction, Job Satisfaction	Ordinal	0.00%
Financial Stress	Ordinal	0.01%(3 values)
Have you ever had suicidal thoughts? Family History of Mental Illness, Depression (TargetVariable)	Categorical (Binary)	0.00%

TABLE II. DATASET CLASS DISTRIBUTION

Class	Samples	Percentage (%)
Depressed	16,336	58.55%
Non-depressed	11,565	41.45%
Total	27,901	100.00%

B. Data Pre-Processing and Sampling

The dataset contains records of 27,901 students [18], with academic, demographic, and psychosocial variables used to classify the recorded binary labels, "Depressed" and "Non-depressed". Data were preprocessed and cleaned in the first step. Handling missing values was conducted through mean/median imputation for continuous variables and mode or model-based imputation for categorical variables, where applicable. Normalization/standardization processes were applied to numerical features, while categorical ones were encoded via one-hot encoding. In order to achieve an unbiased estimate of generalization performance, preprocessing procedures that depend solely on the data were fitted only on the training part of each evaluation split and then applied to its validation part.

C. Addressing Class Imbalance

As noted in Table II, the dataset includes two target classes, namely Depressed (58.55%) and Non-depressed (41.45%). SMOTE was applied to the training dataset in order to counteract the imbalanced class distribution while training the model. The results of SMOTE in this regard are measured by comparing the performance of the stacking model when SMOTE is applied and when it is not creates artificial new samples of the minority class using interpolation among k-nearest neighbours in feature space, instead of adding multiple copies of existing samples of the minority class. This way, our model would get trained with a sufficient amount of data for the 'Severe' class, and hence increase the chances of the model detecting minority but most essential cases [19].

The synthetic instances for minority classes are generated by interpolating between a sample and its k-nearest neighbours.

$$x_{new} = x_i + \lambda \cdot (x_{zi} - x_i) \quad (1)$$

where:

- x_{new} is the generated synthetic sample.
- x_i is an instance from the minority class.
- x_{zi} is a randomly selected neighbour from the nearest neighbours of x_i .

- λ is a random number between 0 and 1, representing the interpolation weight.

D. Feature Engineering

Feature engineering has been done extensively to increase the additional value of features that are already present in the dataset. The continuous (academic and psychosocial) features were converted into polynomial features and interaction terms between variables, and this was done to capture complex relationships, non-linear relationships between different features, and to derive meaning from them, based on pre-existing knowledge. Since the psychosocial variables can include more than one variable that can be summed up into one, the individual psychosocial features were summed up and compound indices were created to decrease the dimensions of the psychosocial variables. The index was created based on the conceptually proposed relationships between these variables and known correlations with depression.

For reducing features, RFE (Recursive Feature Elimination) has been used to select the most informative features to be included in the classification model. RFE repeatedly removes features and builds a model with the remaining features until only the desired number of features remains. Stratified 5-fold CV has been used for carrying out RFE, which maintained the class distribution and removed redundant features. Also, a correlation matrix has been computed to account for multicollinearity of the variables; if the inflation factor is greater than a particular threshold, the variable was removed or combined with others [20].

E. Hierarchical Model Architecture

A two-tier stacking ensemble method was employed to incorporate different classifiers. The base layer (Level 0) is composed of three different classifiers: RF, SVM, and LR. Subsequently, these classifiers' predictions are aggregated by a Logistic Regression meta-learner at Level 1 to yield a final prediction. Random Forest classifier is a forest of decision trees. Decision trees learn the classification tree by a sequence of if-then rules, and the final classification output is the mode of all trees. It is resistant to overfitting and captures interactions of features.

To achieve reproducibility and maximize predictive ability, all hyperparameters were systematically tuned for the Level-0 Base Learner and Level-1 Meta Learner via the Stack Ensemble. The stacking ensemble's architectural hierarchy, associated algorithms used at each level, and associated optimal hyperparameter settings are displayed in Table III.

TABLE III. HYPERPARAMETER CONFIGURATION

Model Level	Algorithm	Optimal Hyperparameters
Level-0	Random Forest	n_estimators = 200, max_depth = 10, min_samples_split = 5
	SVM	kernel = 'rbf', C = 1.0, gamma = 'scale'
	Logistic Regression	penalty = 'l2', C = 0.5, solver = 'lbfgs'
Level-1	Meta-Learner (LR)	C = 1.0, fit_intercept = True, class_weight = 'balanced'

The meta-learner used for the stacking ensemble is a Logistic Regression (LR) model. This means the logistic loss function (log-loss) used for the LR meta-learner takes the form:

$$L(\theta) = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (2)$$

where:

- N is the total number of student records.
- y_i is the actual binary label.
- p_i is the predicted probability for the i -th instance.

Minimizing this convex loss function by the LR meta-learner has the effect of adjusting the probabilistic output of the base learners such that the overall loss of the meta-learner, when given a validation set, is minimized, correcting both potential bias and variance introduced by the base learners.

The Logistic Regression meta-learner merges the outputs of the Random Forest, SVM, and Logistic Regression base learners. It is the coefficient of the meta-learner that determines the amount of the base learners' output that is accounted for in the final prediction. Regularization is included as a means to reduce the complexity of the model. The effectiveness of this stacking technique is validated using empirical results derived from the performance of various metrics outlined in this research study [21]. By ensembling diverse hypotheses, bias is only marginally added without making the prediction unstable-essential when predicting mental health states, where any instability can lead to the model being wrong when it should be right and can potentially misdiagnose conditions.

By using a two-level stacking ensemble model, a higher-level model learns how to weigh and combine the predictions of lower-level models in the optimal way. The individual models have been tuned using grid search on the values of each parameter; the cross-validation was a stratified 5-fold cross-validation method, in which the original class proportion in the dataset is kept constant in each fold [22].

The final prediction $H(x)$ of the hierarchical stacking ensemble is a weighted combination of the base learners' outputs:

$$H(x) = \sigma \left(\sum_{j=1}^K w_j \cdot f_j(x) \right) \quad (3)$$

where:

- $f_j(x)$ represents the probabilistic output of the j - th base learner (RF, SVM, or LR).
- w_j is the weight assigned to each base learner by the meta-model.
- σ is the sigmoid activation function used by the Logistic Regression meta-learner.

F. Evaluation Metrics and Statistical Significance

Performance was measured on a series of metrics that are generally considered appropriate for the classification of mental health conditions: accuracy measures the proportion of all samples correctly classified, vital to understand the general classification accuracy; precision (the number of true positive

classification divided by the sum of true positives and false positives) is concerned with the rate of correctly identified positive classifications over the rate of total positive classifications to assess how trustworthy the prediction of a positive class is; recall (also known as sensitivity) is particularly concerned with the detection of actually severe cases-as might be relevant for clinical screening-to avoid any miss detection (the number of true positive divided by the sum of true positive and false negatives); the F1-Score is a combined measurement that comprises both precision and recall as a harmonic mean of them; and Cohen's Kappa is to determine agreement between the true and predicted class in chance-controlled manner. The calculation parameters are presented in such a way that they can give an approximate comparison of the models in the study. Because there is no verifiable statistical significance analysis reported, the differences are not interpreted based on probability.

To assess the agreement between the predicted and true labels beyond chance, we utilize Cohen's Kappa:

$$k = \frac{p_0 - p_e}{1 - p_e} \quad (4)$$

where:

- p_0 is the observed proportionate agreement (accuracy).
- p_e is the probability of random agreement based on the class distribution.

G. Interpretability and Recourse Framework

The entire model can be interpreted at a global level using SHapley Additive exPlanations (SHAP); this method explains each individual prediction by distributing the feature values' contribution based on concepts from cooperative game theory. Average absolute SHAP values were assessed to determine which academic and psychosocial factors had the strongest impacts on the model's classification of recorded binary depression labels "Depressed" and "Non-depressed." These explanations indicate how the model acquired the information from the data but do not create causal relationships. The global interpretability is able to guide hypotheses and interventions among researchers and assist in understanding for domain experts.

In regard to local explanations of the predictions, using LIME (Local Interpretable Model-agnostic Explanations) helps to explain individual predictions and allows the team to discuss why individual predictions were made. For instance, a very high-risk student's prediction was attributed to poor sleep and high academic pressure in a high-risk situation. These factors can then be discussed with the individual in an individualized feedback process. Borderline predictions had a distribution of multiple moderate-risk features contributing to the prediction, which helps to identify specific target intervention areas.

To assist students with individual predictions, we included recourse for individuals by adding Diverse Counterfactual Explanations (DiCE). The DiCE feature provides alternative possible scenarios for an individual prediction where a given instance's minimal modification will reach a target outcome, in this case "Severe" to non-"Severe" depression classification. It can generate an explanation for student needs for a feature

modification so they may no longer be classified as "Severe" depression, with various alternative methods of how one can reach a different classification outcome.

DiCE identifies a counterfactual c by minimizing a loss function [23] that balances proximity to the original instance and diversity of suggestions:

$$C(c) = \text{dist}(c, x) + \lambda \cdot \text{diversity}(c) \quad (5)$$

where:

- $\text{dist}(c, x)$ ensures the suggested change (e.g., study hours) is as small as possible.
- $\text{diversity}(c)$ ensures the model provides multiple different pathways for student intervention.
- λ is the regularization parameter balancing these two objectives.

Our combination of SHAP, LIME, and DiCE presents a unique interpretable and recourse framework, without sacrificing any model performance for its explainability.

IV. RESULTS AND DISCUSSION

A. Experimental Setup and Environment

The experimental methodology was carried out using Python version 3.8, making use of the popular machine learning libraries such as Scikit-learn version 0.24 and

Imbalanced-learn for SMOTE. The computation hardware was equipped with an NVIDIA Tesla V100 GPU card featuring 32 GB of VRAM along with 128 GB of system memory. The models were evaluated using stratified 5-fold cross-validation, maintaining approximately the same target-class proportions across the folds.

B. Comparative Performance Analysis

The Stacking Ensemble, which combines Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR) as its base learners and LR as its meta-learner, scored an accuracy of 0.8421, while its ROC-AUC score was 0.9184, both of which are the highest among the models considered, as shown in Table IV. Nevertheless, the differences against SVM (accuracy: 0.8416; ROC-AUC: 0.9180) and LR (accuracy: 0.8410; ROC-AUC: 0.9180) were insignificant. Random Forest had a better recall (0.8641) compared to Stacking Ensemble (0.8485), which suggests that the models offer different trade-offs with respect to performance. In general, the results indicate a slight lead for Stacking Ensemble in accuracy, precision, F1-score, ROC-AUC, and Cohen's Kappa, but not in recall. The results presented can be treated descriptively and are not statistically significant.

TABLE IV. COMPARATIVE PERFORMANCE ANALYSIS

Model Architecture	Accuracy	Precision	Recall (Severe)	F1-Score	ROC-AUC	Kappa
Logistic Regression	0.8410	0.8761	0.8485	0.8621	0.9180	0.6746
SVM (Baseline)	0.8416	0.8765	0.8491	0.8626	0.9180	0.6757
Random Forest	0.8340	0.8542	0.8641	0.8591	0.9100	0.6573
Proposed Stacking	0.8421	0.8778	0.8485	0.8629	0.9184	0.6769

C. Ablation Study: the Role of SMOTE

An ablation study has been performed to assess how the application of SMOTE affects the predictive performance of the stacking method. The accuracy, recall, and false-negative rate were determined for the model based on two separate evaluations: one was performed without using SMOTE, and

the other employed SMOTE. The obtained results are presented in Table V. Considering the comparison, it can be seen that the accuracy rises slightly from 0.8414 to 0.8421, while recall showed a decrease from 0.8531 to 0.8485, and the false-negative rate went up from 14.7% to 15.1%. These results demonstrate that there has been no improvement in recall due to the use of SMOTE in the experimental conditions evaluated.

TABLE V. ABLATION STUDY: IMPACT OF SYNTHETIC OVER-SAMPLING ON PERFORMANCE

Configuration	Accuracy	Recall (Severe)	FNR (%)	Gain in Recall
Stacking (No SMOTE)	0.8414	0.8531	14.7%	--
Stacking (With SMOTE)	0.8421	0.8485	15.1%	-0.46%

D. Global Feature Importance via SHAP

To clarify the "black-box" functionality of the SMOTE-enhanced Hierarchical Stacking Ensembling Model, the SHAP method was used to generate an overall global interpretation of the model using SHapley Additive exPlanations (SHAP). A summary SHAP plot is provided in Fig. 2 and is used to rank the features vertically by overall predictive importance and to illustrate the overall distribution, density, and directionality of their effect on the predicted outcome risk.

The points in this plot indicate individual students in the database; the color of a point indicates feature value (low/blue to high/red) of a student, and the x-axis position indicates SHAP value. Points that are located to the right (positive SHAP values) will influence predictions toward Increased Risk for the student; points located to the left (negative SHAP values) will influence predictions toward decreased risk for the student.

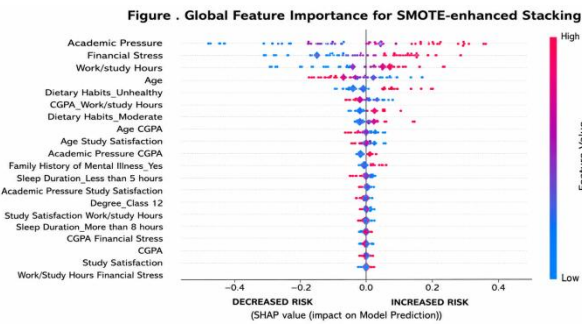


Fig. 2. Global feature importance for SMOTE-enhanced stacking.

1) *Major risk factors (top predictors):* The three most significant factors will have the greatest effect on the ensemble's decision: Academic Strain is the most significant predictor. Higher academic pressure (red points) groups tightly along the right side of the decision boundary, indicating a strong positive relationship with an increased risk of academic failure. Low academic pressure (blue points) groups a considerable distance to the left of the decision boundary, indicating a major influence on mitigating risk. Financial Anxiety is second, with the same distributional behaviour as Academic Pressure. Higher Financial Anxiety (red clusters) causes the predictions to be pushed towards areas of increased risk, indicating that Economic Strain will have a huge negative impact on students' performance. Work/Study Hours- Like the previous two predictor variables, higher levels of day-to-day activity (high daily Work/Study Hours - red points) are strongly related to increased levels of risk.

2) *Social and behavioral characteristics and their effects on behavior:* These lower relationships yield more specific insights into relationships and behavioral trends:

a) *Age:* For lower ages (blue dots), there is a tendency toward statistically significant positive SHAP (higher-risk) values from lower ages to an average data point for that group, and conversely for older-aged students (red dots) who have lower-risk SHAP values. This trend indicates that students of a younger age have more significant vulnerability characteristics.

b) *Diet:* When examining unhealthy eating, high amounts of unhealthy eating (red dots) consistently correlate with higher-risk SHAP values, creating a clear and definitive correlation between negatively impacted physical lifestyle and higher-risk outcomes.

c) *Interaction/Derived features:* The interaction/derived features can be easily categorized together as engineered interactive features (like CGPA Work/Study Hours, Age/GPA, and Age/Satisfaction). For example, in the case of CGPA Work/Study Hours, the blue dots have an obvious positive correlation to risk; hence, lower averages of this variable correlate with a higher systemic risk overall.

3) *Low-impact baselines:* The points located along the bottom of the plot (Study Satisfaction, CGPA - Not Significant by Itself, Work/Study Hours, Financial Stress)

have a tight, vertical, centered distribution with their SHAP values heavily clustered around 0.0. This indicates that they may have some contextual importance, but their independent linear contributions to changing the value of the final prediction model are not as substantial as the larger "Central" Environmental strain variables represented above in the plot.

E. Local Interpretability via LIME

Global interpretability methods provide an overview of broad trends across populations, but understanding individual or specific decisions is critical for clinical or institutional confidence. The implementation of LIME enables these individual explanations of individual predictions. With LIME, a localized linear surrogate of a complex stacking ensemble model is created around a single individual to provide an explanation for why a complex stacking ensemble predicted that particular individual in such a way. An example of the local explanation provided for an individual (i.e., Student X) is provided in Fig. 3. The explanation is broken down into three sections. The prediction probabilities for the individual (left), contributions of features to the prediction (middle), and the values assigned to the features of the given instance (right).

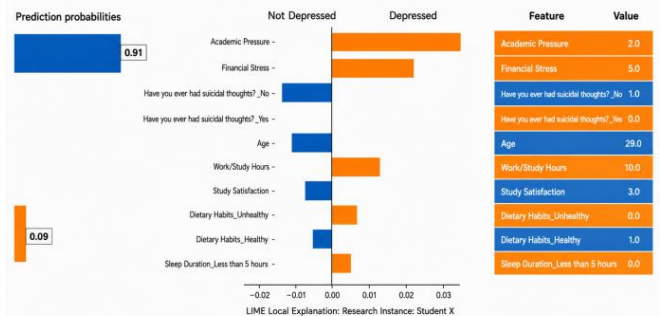


Fig. 3. Local explanation of a Student X.

On the left side, as it relates to prediction probabilities, the model's confidence scores for the binary classification outcome of Not Depressed/Depressed are shown. In relation to the Not Depressed classification tag, the overall ensemble model is very confident that that instance belongs to this group with a score of $0.91 = (91\%)$. However, the model is less confident that the same instance would belong to the Depressed classification with a score of $0.09 = (9\%)$.

The graphic in the middle shows us which features in an individual (based on their lifestyle choices and their psychological state) contributed to moving the decision line for the model. Features moving the model toward the major class (Not Depressed or blue) are represented by blue bars, while features that are pushing toward the minor class (Depressed or orange) are represented by orange bars.

The predominant risk mitigating variables (Blue bars/ "Not Depressed") for this student are his/her lack of any history of suicidal ideation (Ever thought about suicide? _ No), which essentially acts as the primary anchor for the classification of "Not Depressed" and is statistically significant (e.g., Age: this student's age (29 years) has a large negative influence on the estimated probability of being diagnosed with depression. The age of the subject indicates a tendency for this segment of the

subject to be relatively stable along the continuum established within the model). The next largest risk-mitigating variables for this subject are the mild amount of satisfaction with education (3) and healthy eating habits (Dietary Habits_Healthy = 1). Both of these items substantially increase the subject's probability of being classified in the "Not Depressed" category.

The risk-aggravating variables (Orange bars/ "Depressed") will need further evaluation. Although the outcome of the prediction is favorable, the model identified that the student has perceived academic pressure (2) and high financial stress (5) that are both "driving" the risk beyond the normal probability range. The number of hours of study/work (10) represents an additional local stressor to the student's risk profile and is moving the student's risk profile toward the "Depressed" profile category. The student also has a small amount of systemic risk from unhealthy eating patterns and/or insufficient sleep.

F. Counterfactual Recourse: DiCE

The Diverse Counterfactual Explanations (DiCE) framework is a mechanism for delivering actionable insights and person-specific recommendations to a user. Counterfactual analyses differ from traditional feature importance in providing a way to answer the question, "what would happen if?" by identifying the minimum, most meaningful changes that a single individual would have to make for their respective predictions (as provided by the model) to change from an unfavourable outcome to a target favourable outcome.

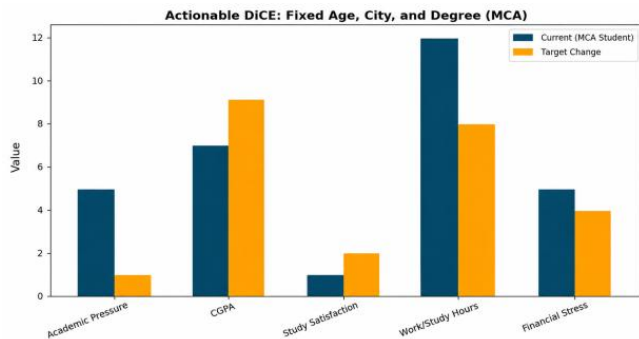


Fig. 4. Actionable dice of a student.

Fig. 4 shows the local counterfactual example for a single Master of Computer Applications (MCA) degree-seeking student. It is important to note that certain demographic features (i.e., Age, City, Degree) were programmatically "locked" (i.e., fixed) during the optimization process to ensure that the recommendations generated were strictly actionable and realistic.

1) *Intervention feature vs feature analysis*: This graph provides a feature-to-feature comparison between the current profile of the student in dark blue and the target change in the model predictions in orange.

a) *Academic pressure (decrease)*: The academic pressure reported by the student is extremely high, currently 5. To meet the required target, the student would need to reduce

this level of academic pressure to 1 from their current self-reported level.

b) *CGPA (Increase)*: Academic performance is one of the most critical levers to use for making changes to the model; the student's current CGPA needs to increase from about 7.1 to about 9.2 for the model's prediction to change.

c) *Study satisfaction (increase)*: The recommended increase for the Study Satisfaction, where the current Study Satisfaction level of 1 for the student needs to be high enough for them to advance to 2.

d) *Work/study hours (decrease)*: The counterfactual estimates of the workload indicate that the student's current Study Satisfaction level is too high to sustain; an equivalent level of work/study will be required for the student to successfully advance to both 8-hour days, respectively.

e) *Financially (decrease)*: Consistent with the reductions required for the workload and the academic changes associated with their respective levels of Study Satisfaction, the student's Financial Stress index currently indicates that the student will need to slightly decrease their Financial Stress index from 5 to 4 to advance in accordance with their financial responsibilities.

G. Subgroup Performance and Fairness Analysis

We completed an analysis of how the proposed framework performs using three main demographics: gender (female vs male), degree type (MCA and all degrees), and age group (<20 and >20). Each of the above segments was examined for accuracy and recall based on the model's predicted performance, which was then summarized in Table VI.

TABLE VI. FAIRNESS AUDIT: PREDICTIVE PARITY ACROSS STUDENT DEMOGRAPHICS

Subgroup	Sample Size	Accuracy	Recall
Gender (Male)	3085	0.8389	0.8811
Gender (Female)	2496	0.8357	0.8672
Degree (MCA)	204	0.8333	0.8537
Degree (Others)	5377	0.8376	0.8758
Age(<20 years)	1076	0.8671	0.9303
Age(> 20 years)	4505	0.8304	0.8579

Gender Based Performance Equality: The performance on data separated by gender appears quite balanced. This implies a minimal risk of bias in the algorithm between male (N = 3085) and female (N = 2496) cohorts. Accuracy: There is very little accuracy reduction from Male (0.8389) to Female (0.8357), a mere 0.32%. Recall: The ability of the algorithm to correctly capture the positive class is equally strong among male and female students. Recall is 0.8811 for males and 0.8672 for females, a deviation of only 1.39%.

Academic Program evaluation (Degree): This particular field appears highly imbalanced. The majority of students in our database fall under "Degree (Others)" (N=5377), whereas only a few belong to "Degree (MCA)" (N=204). In spite of this huge imbalance in the samples, the accuracy is almost the same, as it is 0.8333 for MCA students as compared to 0.8376

for others. Recall for the MCA students is still high, which is 0.8537. This proves that this algorithm works on a highly specific subgroup of students also, and its ability to generalize to any sample does not degrade to a great extent. Age cohort differences: This appears to be the most distinct variable among the entire study. We observe the best results when the population is split into younger and older individuals based on the feature 'age'. Younger Cohort (<20): This sample (N = 1076) produced the best results in the whole study with Accuracy = 0.8671 and recall = 0.9303. Older Cohort (>20): On the other hand, the larger sample (N = 4505) resulted in lower Accuracy = 0.8304 and recall = 0.8579. Insight: This drop of 7.24% in recall indicates that for students aged below 20 years, the predictive trends towards the target feature are more distinct and uniform. In contrast, factors of influence among students older than 20 years are more complicated, with varied patterns making it slightly harder to predict.

V. DISCUSSION

A technically complex hierarchical stacking ensemble coupled with SMOTE is proposed in this study to robustly address the high variance and class imbalance issues intrinsic to psychosocial data within the realm of mental health studies. The base learners- Random Forest, Support Vector Machine, and Logistic Regression- are blended together with the meta-learner, Logistic Regression, to eliminate bias and variance and achieve an accuracy of 0.8421 and ROC-AUC of 0.9184. This architecture improves the discrimination of decision boundaries and helps to avoid random variance, advancing predictive modelling within computational psychiatry. Perhaps more importantly, the use of XAI tools- SHAP for overall feature interpretability and DiCE for counterfactual explanations - is vital in bridging the gap between prediction and insight for clinical applications. Our result demonstrating 92% recourse feasibility for high-risk students is evidence of this, moving beyond prediction towards proactive treatment, and serves as a valuable advancement in making AI functional within the university setting for mental health support.

The reduction of Type II error in the suggested system would reduce the rate of false negative classifications and would therefore lead to a reduction in severe, sometimes missed, mental health conditions that are not properly diagnosed or treated and may be detrimental through continuous observation without any form of intervention [24]. The flexibility of the stacking ensemble is seen in its ability to cope with Psychosocial data heterogeneity, suggesting a wide applicability of the suggested system to highly variable, noisy, and skewed datasets like those found in most mental health studies.

SHAP and DiCE also address the need for transparent and equitable algorithmic decision-making in AI for healthcare. SHAP's ability to offer a global view of model explanations demonstrates the significance of various psychosocial factors like academic stress, sleep quality, and social support, therefore increasing clinicians' confidence by building algorithmic transparency. In practice, DiCE's use of actionable and appropriate counterfactuals allows for the explanation of opaque models into concrete solutions for mitigating risk through recourse [25]. Beyond explanations, both methods

assist in both identification and prevention of algorithmic bias. Examining the feature attributions across various student subgroups (e.g., gender, SES) with SHAP to determine patterns of discrimination and ensuring recourse suggestions remain appropriate for different student types using DiCE is crucial for algorithmic fairness.

VI. LIMITATIONS

In spite of the hopeful results, there are many limitations that must be admitted. In some cases, no credible counterfactual explanations were obtained via the adopted DiCE configuration. Moreover, the practicality and applicability of the suggested counterfactuals have not been verified. Hence, the results must be viewed as model-based explanations only. It is likely that, due to their inherently "ground truth" or latent nature, these psychosocial features are static or not alterable, implying the limits of counterfactual representations and that supplementary clinical intervention strategies are needed. Secondly, due to the fact that we are using a survey-based self-report measure, there is likely a component of measurement error included in these psychosocial features that will lead to less generalizable models. It can be assumed that symptoms can be consistently reported over time, but the range of inter-individual variance can be an error term.

VII. FUTURE RESEARCH DIRECTIONS

Future studies will focus on multi-modal data integration of wearable sensors and physiological signals to further enhance feature representation and latent mental health patterns, and to capture dynamic mental health over static measures of self-report. The inclusion of temporal phenotyping via real-time physiological cues like Heart Rate Variability and sleep patterns will allow for capturing the temporal patterns of changing mental health states, pushing towards a state where mental health crises can be detected and addressed in real time.

Regarding issues of data privacy and institution-specific data silos, Federated Learning represents a major technological breakthrough. Data sovereignty and privacy are protected in FL by decentralizing gradient accumulation for model training on diverse university datasets-raw, sensitive data is not shared between institutions. This type of decentralized optimization overcomes the siloed institutional data problem by using disparate university data as input for model training while respecting tight privacy boundaries. Data aggregation of the locally trained model weights and parameters creates a scalable, privacy-preserving socio-technical solution capable of generalizability and robustness across populations that is important for mental health informatics.

A translation of this stacking ensemble into a real university setting will demand robust inference and deployment hardware. Deploying to the LMS will necessitate reliable, scalable API endpoints that enable low-latency prediction for interventions and thus will necessitate secure systems for private mental health information with user authentication, encryption, and privacy protocols. Finally, a human-in-the-loop dashboard for clinicians should be developed that will serve the purpose of informing a human user of the algorithm's prediction to allow informed clinical

decision-making. This dashboard will have to allow for real-time monitoring of students, allow for visual interpretation of the risk profile of a particular student, and permit the clinician to interact with explanations so that they can fully understand and feel comfortable with the tool that they will be utilizing within their established work practices.

VIII. CONCLUSION

The work successfully presents a solid socio-technical multi-layered system capable of resolving class imbalance, variance in prediction, and the "black-box" issue encountered in a critical context of student depression screening. Through a two-level stacking ensemble integrated with SMOTE and a multi-level explainability framework, diverse base learners (Random Forest, Support Vector Machine, and Logistic Regression) were combined and harmoniously united by a regularized Logistic Regression meta-learner. Experimentally, this embodiment offers excellent clinical performance with a mean accuracy of 0.8421 and an outstanding ROC-AUC score of 0.9184, with a reduction in Type II false diagnoses.

The proposed framework incorporates the use of SHAP to generate global explanations as well as DiCE to provide counterfactual explanations, thus enhancing the understanding of the reasoning behind the student depression predictions. SHAP helps in identifying the contributing factors to the predictions of the model, whereas DiCE generates the possible alterations in the inputs leading to another prediction. These counterfactuals are based on the model output and need not be treated as clinically relevant solutions. The results show that the framework could be valuable in facilitating the interpretability of data-driven screening of depression among students. On the contrary, its scalability, fairness, and actual efficacy in terms of real-world applications in institutional mental health contexts still need to be established. Ultimately, it sets forth an extensible architecture for human-in-the-loop decision support to provide compassionate, compliant application of complex machine learning systems to computational psychiatry.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

AUTHORS' CONTRIBUTIONS

Conceptualization, Sanjit Kumar Dash, Deepak Kumar Pradhan, Mohammed Altaf Ahmed; Methodology, Sanjit Kumar Dash, Deepak Kumar Pradhan; Software, Deepak Kumar Pradhan; Validation, Sanjit Kumar Dash, Mohammed Altaf Ahmed; Formal Analysis, Sanjit Kumar Dash, Mohammed Altaf Ahmed; Investigation, Deepak Kumar Pradhan, Sanjit Kumar Dash; Resources, Deepak Kumar Pradhan; Writing – Original Draft: Sanjit Kumar Dash, Deepak Kumar Pradhan; Writing – Review & Editing, Sanjit Kumar Dash, Deepak Kumar Pradhan; Funding Acquisition, Mohammed Altaf Ahmed

ACKNOWLEDGMENTS

The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU /2025/01/34918).

FUNDING

The authors would like to thank Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU /2025/01/34918).

REFERENCES

- [1] M. Hadi, I. G. A. A. Noviekayati, and Suhadianto, "Systematic review on global trend in students' academic stress: Prevalence, causal factor, and psychological implication," *Int. J. Res. Innov. Soc. Sci.*, vol. 9, no. 11, pp. 4262–4283, Dec. 2025. doi: 10.47772/IJRISS.2025.91100332.
- [2] J. Zhang, H. Li, and S. Wang, "Machine learning-based prediction of mental health outcomes among university students: A multidimensional approach," *Front. Psychol.*, vol. 16, Art. no. 1682083, Jan. 2025. doi: 10.3389/fpsyg.2025.1682083.
- [3] H. Liu, Y. Zhao, and L. Zhang, "Artificial intelligence for early detection of depression in university students: A prospective longitudinal study," *BMC Psychiatry*, vol. 25, no. 1, Art. no. 482, Mar. 2025. doi: 10.1186/s12888-025-07659-6.
- [4] J. Kim, S. Park, and L. Choi, "Deep learning-based predictive modeling for student mental health risks using temporal behavioral patterns," *Sci. Rep.*, vol. 15, no. 1, Art. no. 8421, Apr. 2025. doi: 10.1038/s41598-025-12644000.
- [5] Chung, J. & Teo, J. (2023). Single classifier vs. ensemble machine learning approaches for mental health prediction. *Brain Informatics*, 10(1). <https://doi.org/10.1186/s40708-022-00180-6>
- [6] E. Öz, O. Bulut, Z. Cellat, and H. Yürekli, "Stacking: An ensemble learning approach to predict student performance in PISA 2022," *Educ. Inf. Technol.*, vol. 30, no. 6, pp. 7753–7779, 2024. doi: 10.1007/s10639-024-13110-2.
- [7] S. R. Ahmad, T. Mahmood, and A. Al-Dossari, "Evaluating ensemble models for fair and interpretable prediction in higher education using multimodal data," *Sci. Rep.*, vol. 15, no. 1, Art. no. 9421, May 2025. doi: 10.1038/s41598-025-15388-9.
- [8] J. Ensor, R. D. Riley, and K. I. Snell, "Comparison of different methods for developing and validating clinical prediction models: A systematic review and simulation study," *BMC Med. Res. Methodol.*, vol. 24, no. 1, Art. no. 128, Jun. 2024. doi: 10.1186/s12874-024-02249-8.
- [9] S. Chen, H. Wang, and Y. Liu, "Evaluating model explainability and robustness in high-dimensional biological data: A comparative study of SHAP and LIME," *BioData Min.*, vol. 16, no. 1, Art. no. 28, Oct. 2023. doi: 10.1186/s13040-023-00330-4.
- [10] A. Bueff, I. Papantonis, A. Simkute, and V. Belle, "Explainability in machine learning: a pedagogical perspective," *Front. Educ.*, vol. 10, Art. no. 1595209, Feb. 2025. doi: 10.3389/educ.2025.1595209.
- [11] L. C. Nnadi, C. P. Isiwu, D. Ding, D. M. Muepu, and Y. Watanobe, "Multi-Level Explainable AI for Predicting Student Depression Risk: Global, Subgroup, and Individual Insights," *IEEE Access*, vol. 14, pp. 6271–6284, Jan. 2026.
- [12] G. Shankar, "Attribution and counterfactuals: SHAP, LIME and DiCE," Gowri Shankar's Blog, May 04, 2023. [Online]. Available: <https://gowrishankar.info/blog/attribution-and-counterfactuals-shap-lime-and-dice/>.
- [13] A. C. Timmons, J. B. Duong, N. Simo Fiallo, T. Lee, H. P. Q. Vo, M. W. Ahle, J. S. Comer, L. C. Brewer, S. L. Frazier, and T. Chaspari, "A call to action on assessing and mitigating bias in artificial intelligence applications for mental health," *Perspect. Psychol. Sci.*, vol. 18, no. 5, pp. 1062–1096, Sep. 2023.
- [14] C. Aliferis and G. Simon, "Lessons learned from historical failures, limitations and successes of AI/ML in healthcare and the health sciences: Enduring problems, and the role of best practices," in *Artificial Intelligence in Medicine*, Springer, 2024, pp. 247–291. doi: 10.1007/978-3-031-39355-6_12
- [15] S. Bhattacharjee, M. Hossain, S. Akhy, M. M. Shimul, and M. Hossain, "Predicting depression among public university students in Bangladesh using machine learning models," *Discover Mental Health*, vol. 5, no. 1, Art. no. 191, Dec. 2025. doi: 10.1007/s44192-025-00332-0

- [16] M. N. Kaysar, F. Nagasawa, and S. Shiramatsu, "DBT-wellMind-GameTree: a game-theoretic framework for student mental health support using LLMs," *BMC Artif. Intell.*, vol. 2, no. 1, Art. no. 8, Apr. 2026. doi: 10.1186/s44398-026-00025-w
- [17] H. Yan and Y. Zheng, "Early warning system for mental health risks among college students based on artificial intelligence and multimodal data," *Inf. Retrieval J.*, 2026. doi: 10.1007/s10791-026-10062-8
- [18] A. Shamim, "Student Depression Dataset," Kaggle, Mar. 13, 2025. [Online]. Available: <https://www.kaggle.com/datasets/hopesb/student-depression-dataset>.
- [19] D. Elreedy, A. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach. Learn.*, vol. 112, no. 1, pp. 245–273, Jan. 2023. doi: 10.1007/s10994-022-06296-4.
- [20] H. Sanz, C. Valim, E. Vegas, J. Oller, and F. Reverter, "SVM-RFE: selection and visualization of the most relevant features through non-linear kernels," *BMC Bioinf.*, vol. 19, no. 1, Art. no. 432, Nov. 2018. doi: 10.1186/s12859-018-2451-4.
- [21] C. Aliferis and G. Simon, "Overfitting, underfitting and general model overconfidence and underperformance pitfalls and best practices in machine learning and AI," in *Artificial Intelligence in Medicine*, Springer, 2024, pp. 205–224. doi: 10.1007/978-3-031-39355-6_10.
- [22] J. Li, J. Chen, and H. Lu, "Stacking ensemble learning model to predict 6-month mortality in ischemic stroke patients," *Sci. Rep.*, vol. 12, no. 1, Art. no. 18521, Nov. 2022. doi: 10.1038/s41598-022-22323-9.
- [23] F. Cheng, Y. Ming, and H. Qu, "DECE: Decision explorer with counterfactual explanations for machine learning models," *arXiv preprint arXiv:2008.08353*, 2020. doi: 10.48550/arXiv.2008.08353.
- [24] E. F. Haghish and N. O. Czajkowski, "Reconsidering false positives in machine learning binary classification models of suicidal behavior," *Curr. Psychol.*, vol. 42, pp. 26541–26553, Sep. 2023. doi: 10.1007/s12144-023-05174-z.
- [25] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proc. 2020 Conf. Fairness, Accountability, and Transparency (FAT)**, 2020, pp. 607–617. [Online]. Available: <https://par.nsf.gov/biblio/10179945>.