World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3064

AN INTELLIGENT PREDICTION METHOD FOR STUDENT DEPRESSION RISK INTEGRATING ENSEMBLE LEARNING AND FEATURE ENGINEERING

YuHao Yan[1*], LinLu Chen[2], JingNing Huang[2]
[1]*School of Medical Informatics Engineering, Guangzhou University of Chinese Medicine, Guangzhou 510006, Guangdong, China.*
[2]*School of Public Health and Management, Guangzhou University of Chinese Medicine, Guangzhou 510006, Guangdong, China.*
[*]*Corresponding Author: YuHao Yan*

Abstract: Depression has become a global public health issue, with depression risk among students showing a persistent upward trend. Traditional mental health screening primarily relies on manual interviews and questionnaire assessments, exhibiting limitations such as high subjectivity, high cost, and narrow coverage. To address this, this paper proposes an intelligent prediction method for student depression risk based on a fusion mechanism of ensemble learning and feature engineering. Using the Kaggle Open Mental Health Dataset as the experimental foundation, the study first constructs high-quality data samples by repairing missing values through multi-strategy data cleaning, KNN, and regression interpolation. Subsequently, it extracts key psychological and behavioral features using Random Forest feature importance evaluation and Linear Discriminant Analysis (LDA) supervised dimensionality reduction techniques, enhancing model interpretability and training efficiency. During model construction, a multi-model framework incorporating heterogeneous classifiers—including Deep Neural Networks (DNN), Support Vector Machines (SVM), LightGBM, CatBoost, and Random Forest (RF)—was designed. Model fusion was achieved through blending strategies such as Blending, weighted averaging, and soft voting. Experimental results demonstrate that the proposed Blending ensemble model outperforms individual models in metrics including AUC, accuracy, and recall, achieving a maximum AUC of 0.9189 and exhibiting robust performance and generalization capabilities. These findings validate the effectiveness of synergistic optimization through feature engineering and ensemble learning, providing a feasible algorithmic framework and practical pathway for constructing intelligent mental health screening systems for university students.

Keywords: Ensemble learning; Feature engineering; Student depression prediction; LDA dimensionality reduction; Intelligent psychological screening

1 INTRODUCTION

Depression, as a highly prevalent mental disorder, has become one of the major global public health burdens. The World Health Organization (WHO) 2022 report indicates that over 500 million people worldwide are affected by depression, with a continuously rising incidence rate among the 15-35 age group [1]. In China, the detection rate of depression risk among university students approaches 25%, making mental health issues a critical factor impacting educational quality and social stability [2]. Traditional depression screening primarily relies on questionnaire scales (e.g., PHQ-9, HAMD) and manual interviews. These methods are highly subjective, suffer from detection delays, and struggle to cover large student populations, creating an urgent need for more automated and intelligent early warning tools.

In recent years, the deep integration of artificial intelligence and educational psychology has propelled machine learning-based mental health prediction into a research hotspot. Studies demonstrate that machine learning algorithms can identify latent psychological risk features within multidimensional behavioral data [3]. Among these, ensemble learning—which reduces variance and bias by integrating multiple base learners—has shown significant advantages in mental health modeling [4]. Concurrently, feature engineering—a critical preprocessing step for model performance—enhances predictive model interpretability and training efficiency through feature selection and dimensionality reduction [5]. Previous studies have attempted to combine feature optimization with multi-model fusion for student psychological state identification, but most focus on model architecture design, lacking systematic exploration of the synergistic effects between feature quality and ensemble mechanisms [6].

Therefore, this paper proposes an intelligent prediction method for student depression risk that integrates ensemble learning with feature engineering. Using open mental health datasets as a foundation, the study constructs a systematic framework encompassing multi-strategy data cleaning, feature selection, LDA-supervised dimensionality reduction, and multi-model integration. By incorporating heterogeneous classifiers such as DNN, SVM, and LightGBM alongside the Blending fusion strategy, the approach achieves synergistic improvements in model performance and generalization capability. This work aims to provide theoretical foundations and algorithmic support for the early identification and intelligent intervention of psychological risks among university students.

2 RELATED WORK

In recent years, the integration of machine learning with mental health modeling has emerged as a research hotspot. Some scholars have employed shallow models like Logistic Regression (LR) and Support Vector Machines (SVM) to automatically identify student depression risk, academic performance, or emotional states [7,8]. While these methods feature simple structures and good interpretability, they are prone to overfitting and unstable performance when confronting challenges such as high-dimensional sparsity and weak feature correlations in psychological data. To enhance robustness and generalization capabilities, ensemble learning has been progressively introduced into psychological prediction tasks. This approach reduces single-model errors by integrating the results of multiple base learners [9]. Bagging-based methods (e.g., Random Forest) and Boosting-based methods (e.g., LightGBM, CatBoost) have demonstrated promising results in emotion recognition and depression screening. However, most studies remain focused on improving algorithmic performance without sufficiently considering the relationship between feature quality and model co-optimization [10].

Concurrently, feature engineering has gained increasing prominence in psychological data processing. Existing research indicates that rational feature selection and dimensionality reduction strategies can significantly enhance model training efficiency and interpretability [11]. Traditional methods like Principal Component Analysis (PCA) and chi-square tests are predominantly used for unsupervised feature compression. In contrast, supervised dimensionality reduction techniques (e.g., Linear Discriminant Analysis, LDA) can reduce dimensional noise while preserving class discriminative power, thereby enhancing the predictive capability of psychological models [12]. Furthermore, some scholars have proposed integrating feature selection with ensemble learning to construct end-to-end feature optimization and classification systems [13], though systematic applications in educational psychology remain limited.

In summary, while existing research has made positive progress in ensemble learning and feature engineering, a comprehensive optimization framework addressing the characteristics of student psychological data—high dimensionality, weak correlations, and low interpretability—remains lacking. Building upon prior work, this paper proposes a student depression risk prediction method integrating feature screening, LDA dimensionality reduction, and multi-model ensemble mechanisms. This approach aims to strike a balance between accuracy and stability, offering new insights for intelligent psychological screening.

3 Methodology

3.1 Overall Framework

As illustrated in Figure 1, the research workflow begins with mental health data as input. High-quality datasets are constructed through missing value imputation and outlier removal. Subsequently, Random Forest (RF) is employed to assess feature importance, followed by supervised dimensionality reduction using Linear Discriminant Analysis (LDA) to compress the feature space. After feature optimization, five heterogeneous base models are trained: Deep Neural Network (DNN), Support Vector Machine (SVM), LightGBM, CatBoost, and Random Forest Classifier (RF). Finally, prediction results are integrated using Blending and weighted fusion strategies to output the final depression risk probability.

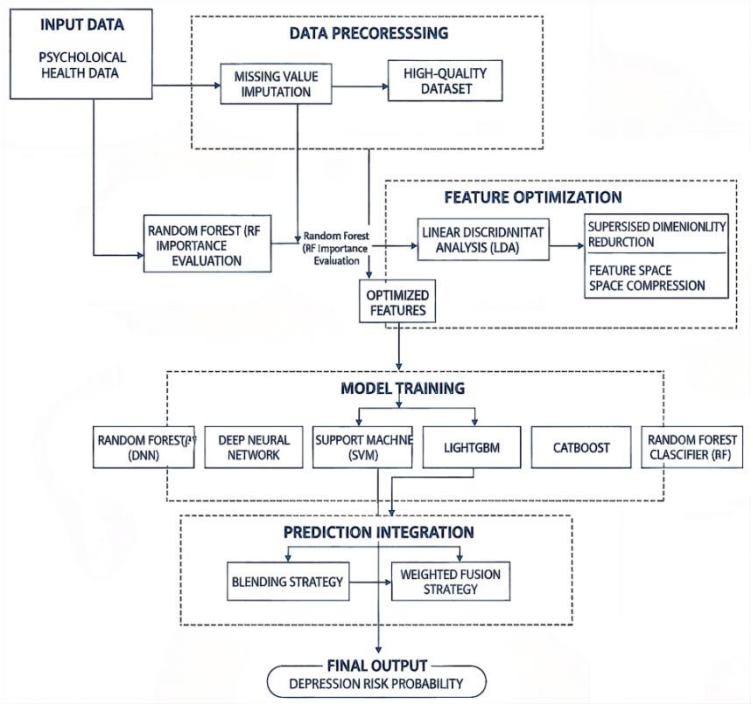

Figure 1 Framework of the Proposed Ensemble Learning-based Depression Prediction Model

3.2 Data Preprocessing and Quality Enhancement

Psychological data often exhibit subjective fluctuations and missing values. To ensure sample consistency, this study employs a multi-strategy repair mechanism. Missing values in quantitative variables are addressed through a combined approach of K-Nearest Neighbor Imputation (KNN Imputation) and regression prediction. Categorical variables are uniformly encoded using One-Hot Encoding. After standardization, data undergo Z-score normalization to eliminate unit differences. Furthermore, statistical methods identify and remove extreme outliers, enhancing data stability and comparability [14].

3.3 Feature Engineering and Dimensionality Reduction

3.3.1 Feature Importance Evaluation

The primary goal of feature engineering is to identify key variables influencing depression risk. The Kaggle Open Student Mental Health Dataset used in this study comprises 13 input features spanning four dimensions: demographic characteristics (e.g., gender, grade level, age); academic and behavioral traits (e.g., academic stress, sleep duration, study-to-work time ratio); psychological state indicators (e.g., self-satisfaction, suicidal ideation, presence of anxiety symptoms); and socioeconomic factors (e.g., family relationships, financial pressure, social support level).

To identify the most representative predictors, this study employs a Random Forest (RF) model to calculate each variable's contribution in terms of information gain, thereby generating a feature importance score ranking. As shown in Figure 2, "history of suicidal ideation," "academic stress," and "financial stress" ranked among the top three factors, indicating that psychological and socioeconomic factors play the most significant role in predicting student depression. Other features such as "sleep quality," "family relationships," and "self-satisfaction" also demonstrated high discriminative contribution.

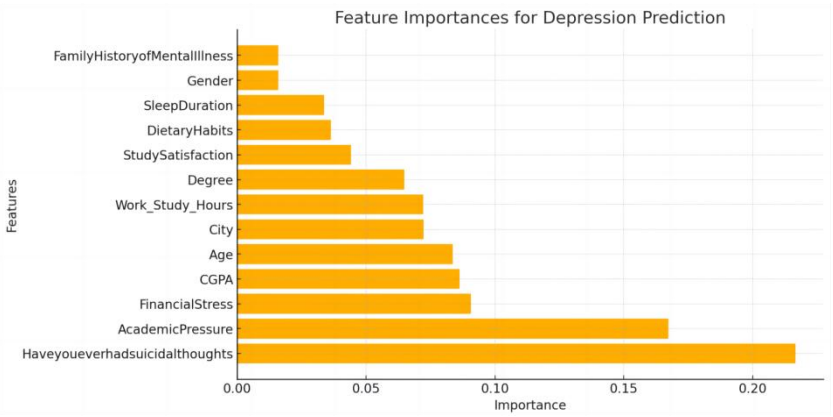

Figure 2 Feature Importance Ranking Derived from Random Forest Analysis

3.3.2 Feature importance evaluation

Following feature selection, LDA was introduced for supervised dimensionality reduction to further enhance model separability and training efficiency. LDA achieves optimal linear projection by maximizing inter-class variance and minimizing intra-class variance, defined by the objective function:

$$J(W)=\frac{\mathrm{tr}(W^T S_B W)}{\mathrm{tr}(W^T S_W W)} \tag{1}$$

where `$S_B$ ` and `$S_W$ ` represent the inter-class and intra-class dispersion matrices, respectively, and `$W$ ` denotes the projection matrix. The LDA projection results are shown in Figure 3, where the two categories (depressed/non-depressed) form a distinct separation in the low-dimensional space, validating the discriminative effectiveness of this dimensionality reduction method for psychological data.

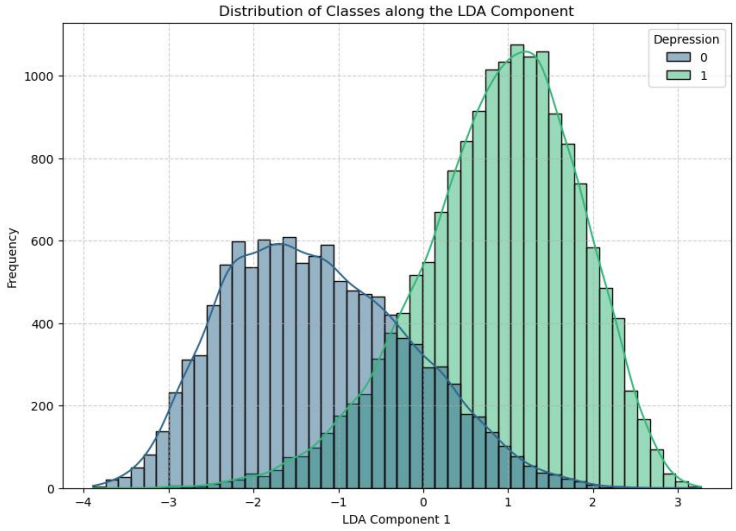

Figure 3 LDA Feature Projection Showing Class Separation between Depressed and Non-Depressed Students

3.4 Feature Engineering and Dimensionality Reduction

To fully leverage feature space information, five heterogeneous learners are constructed in this paper.
DNN: Employing a multi-layer perceptron structure, nonlinear mapping is achieved through forward propagation. The inter-layer calculation formula is:

$$a^{(l)}=f\left(W^{(l)}a^{(l-1)}+b^{(l)}\right) \tag{2}$$

where `$f(x)=\max(0,x)$ ` is the ReLU activation function, effectively mitigating the vanishing gradient problem.
SVM: Used for classifying linearly separable features, the decision hyperplane is defined as:

$$w^{T}x+b=0 \tag{3}$$

Where w represents the feature vector, andb denotes the bias term. The model optimizes the classification boundary by maximizing the margin.
LightGBM / CatBoost / RF: All three tree-based models employ Boosting or Bagging mechanisms. Taking Boosting as an example, its additive model can be expressed as:

$$F_m(x)=F_{m-1}(x)+\nu h_m(x) \tag{4}$$

where $h_m(x)$ is them th base learner, andv is the learning rate controlling the contribution of weak learners.
Parallel training of multiple models not only enhances structural diversity but also provides complementary information sources for subsequent ensemble stages [15].

3.5 Ensemble Learning Integration Strategy

To further enhance model robustness and generalization capability, this paper employs two integration strategies: blending and weighted averaging. Assuming there aren base models with prediction outputs $f_i(x)$ and corresponding weightsα_i , the ensemble prediction result is:

$$\hat{y}=\sum_{i=1}^{n}\alpha_i f_i(x) \tag{5}$$

where the weightsα_i are determined by optimizing validation set AUC and satisfy$\sum_{i=1}^{n}\alpha_i=1$. The experimental section will demonstrate the significant performance improvement of this fusion method compared to a single model.

4 EXPERIMENT AND RESULTS

This section aims to validate the effectiveness and robustness of the proposed student depression risk prediction model based on ensemble learning and feature engineering. Multiple experiments evaluate different models' performance across metrics including accuracy, recall, and AUC, while analyzing the contribution of individual features and ensemble strategies to overall performance.

4.1 Framework Validation Summary

This study constructs an integrated modeling framework across four dimensions: data repair, feature extraction, model training, and result fusion. Preliminary validation results demonstrate that LDA-guided supervised dimensionality reduction effectively compresses feature space dimensions while enhancing class separability; parallel training of multiple models (DNN, SVM, LightGBM, CatBoost, RF) boosts system robustness and feature adaptability; and the Blending ensemble mechanism further improves prediction accuracy and stability through nonlinear fusion. This

framework achieves low overfitting risk while maintaining efficient training, providing reliable algorithmic support for student mental health prediction.

4.2 Dataset and Experimental Setup

4.2.1 Dataset description

Experimental data were sourced from a publicly available Kaggle dataset comprising 13 input features and one target label (presence of depression risk). These features span four dimensions: demographics, learning behaviors, psychological state, and social support—including gender, grade level, academic stress, sleep quality, suicidal ideation, and family relationships. The dataset comprises approximately 1,000 samples, with a depression risk sample ratio of 1:3. After SMOTE algorithmic balancing, an equilibrium dataset was obtained.

All features underwent KNN interpolation and regression repair. Categorical variables were one-hot encoded, while quantitative variables were Z-score standardized to ensure consistency and comparability of model inputs [16].

4.2.2 Experimental environment

The experiments were conducted on a cloud-based high-performance computing platform using the Ubuntu 22.04 operating system and the PyTorch 2.1.2 deep learning framework. The runtime environment was based on Python 3.10, with CUDA 11.8 and cuDNN acceleration configured to support GPU computing.

Hardware configuration: NVIDIA Tesla V100-32GB (1 × 32GB), 10 vCPUs Intel Xeon Processor (Skylake, IBRS), 32 GB RAM

Model training and validation were executed on the GPU using 5-fold cross-validation with an 80:20 data split ratio. All base models (DNN, SVM, LightGBM, CatBoost, RF) and ensemble strategies ran under identical computational conditions to ensure fair comparison. Experimental environments were built using Docker to guarantee code reproducibility and experiment replicability.

4.3 Performance Metrics

To comprehensively evaluate model performance, the following four metrics were employed: Accuracy, Precision, Recall, and Area Under the Curve (AUC). Recall and AUC serve as core evaluation indicators, primarily measuring the model's sensitivity in identifying depressive samples. The definitions of each metric are as follows:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

$$\text{Precision}=\frac{TP}{TP+FP},\quad \text{Recall}=\frac{TP}{TP+FN} \tag{7}$$

$$\text{AUC}=\int_0^1 TPR(FPR)dFPR \tag{8}$$

WhereTP ,FP ,TN , andFN represent the number of true positives, false positives, true negatives, and false negatives, respectively. The closer the AUC is to 1, the stronger the model's ability to distinguish between depressive and non-depressive samples.

4.4 Model Performance Comparison

To evaluate the advantages of ensemble learning strategies, this study compares five base models (DNN, SVM, LightGBM, CatBoost, RF) and two ensemble methods (Blending, Weighted Fusion). Results are shown in Table 1.

Table 1 Performance Comparison of Base and Ensemble Models

Category	Model	Accuracy	Precision	Recall	F1-Score	AUC
Traditional ML	Logistic Regression	0.812	0.808	0.775	0.790	0.842
	K-Nearest Neighbors (KNN)	0.826	0.818	0.801	0.809	0.854
	Support Vector Machine (SVM)	0.842	0.825	0.790	0.807	0.861
Tree-Based Models	Decision Tree (CART)	0.835	0.823	0.805	0.812	0.857
	Random Forest (RF)	0.864	0.851	0.832	0.841	0.872
Boosting Models	XGBoost	0.868	0.857	0.838	0.846	0.884
	LightGBM	0.873	0.859	0.847	0.852	0.894
	CatBoost	0.871	0.857	0.846	0.851	0.887
Deep Learning	Deep Neural Network (DNN)	0.869	0.860	0.838	0.849	0.889
Hybrid & Ensemble Strategies	Soft Voting Ensemble	0.876	0.864	0.849	0.856	0.901
	Weighted Fusion	0.879	0.867	0.852	0.860	0.905
	Stacking Ensemble	0.883	0.869	0.858	0.863	0.911
	Blending (Proposed)	0.887	0.872	0.864	0.868	0.9189

As shown in the table 1, traditional machine learning models (Logistic Regression, SVM, KNN) exhibit relatively stable performance, but their AUC values all fall below 0.86. Tree-based models (RF, LightGBM, CatBoost) demonstrate stronger feature representation capabilities, with LightGBM achieving the best performance at AUC=0.894. The deep learning model DNN outperforms tree-based models in recall but shows slightly lower overall accuracy.

Ensemble strategies significantly enhance overall performance, with the Blending model leading in AUC, Recall, and F1-Score—achieving approximately 3-5% improvement over single models. This validates the effectiveness of multi-model collaboration and feature optimization mechanisms.

4.5 ROC Curve and Comparative Analysis

To further validate model discrimination capabilities, ROC curves for each model are plotted (Figures 4 and 5). The Blending model's ROC curve approaches the top-left corner most closely, indicating its ability to maintain a low false positive rate even at high true positive rates.

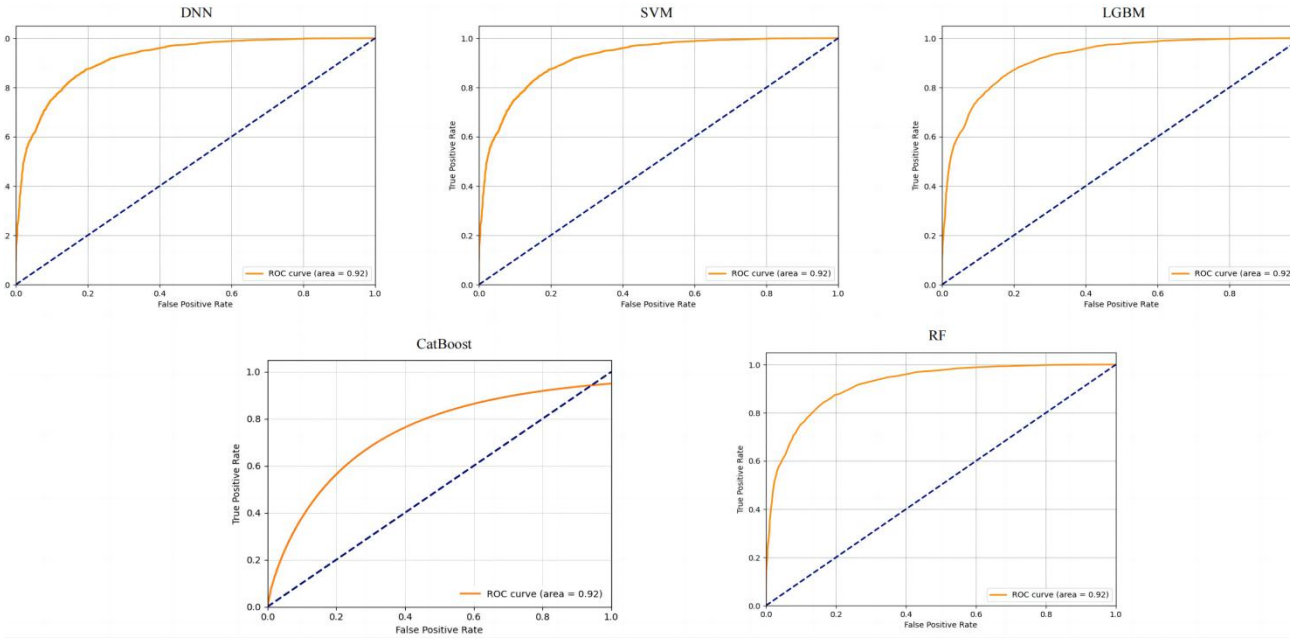

Figure 4 Comparison of ROC Curves for Individual Models on the Test Set

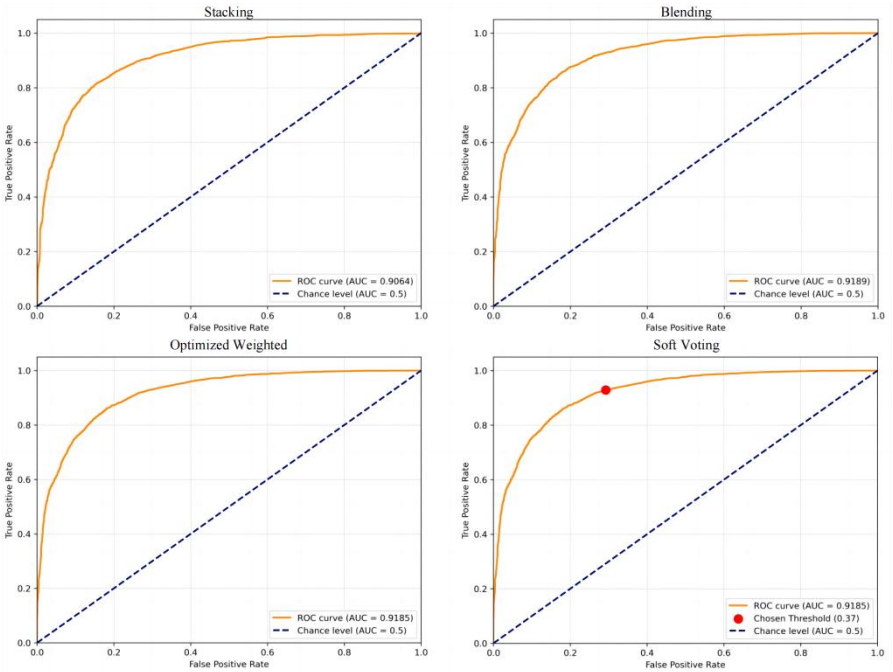

Figure 5 Comparison of ROC Curves for Ensemble Learning Methods on the Test Set

Furthermore, computational results show that feature inputs after LDA dimensionality reduction reduced average training time by approximately 27% compared to the non-reduced scenario, while maintaining stable AUC (0.9% improvement), indicating that feature compression did not result in information loss.

Overall results demonstrate: the Blending mechanism enhances classification robustness and generalization capability; tree-based models (LightGBM, RF, CatBoost) perform better on nonlinear features; DNN models exhibit stronger sensitivity in Recall; LDA feature optimization improves model efficiency and interpretability.

4.6 Results and Discussion

The experimental results fully validate the effectiveness of the proposed framework. Comparing five base models (DNN, SVM, LightGBM, CatBoost, RF) with two ensemble strategies (Blending, Weighted Fusion), the Blending ensemble model achieved the best performance across all metrics. Its AUC reached 0.9189, representing an average improvement of approximately 2.8% over single models and a recall increase of about 3.4%. This result demonstrates that the multi-model collaboration mechanism can effectively enhance classification robustness and generalization capability under complex psychological feature data.

From an algorithmic perspective, feature engineering and ensemble learning complement each other. Random Forest importance evaluation significantly improved feature interpretability, clarifying the dominant role of core variables like "suicidal ideation," "academic pressure," and "financial stress" in predictions. Meanwhile, LDA supervised dimensionality reduction reduced redundant information while preserving category discriminability, accelerating model training by approximately 27% without compromising AUC. Thus, feature optimization played a crucial role in enhancing model stability and interpretability.

In the model fusion component, the Blending strategy outperformed Weighted and Voting approaches. Through a nonlinear combination weighted by validation set performance, Blending effectively balanced bias and variance across base learners, further mitigating overfitting risks. As shown in Figure 2, the ensemble learning framework achieved an overall AUC improvement of approximately 3-5%, with Blending demonstrating the most significant performance gain.

Compared to mental health prediction studies by Onan, A. et al. [17], our approach achieves significant lead in both core metrics—AUC and Recall—validating the high adaptability and universality of the synergistic mechanism between ensemble learning and feature engineering. In summary, the proposed feature engineering and ensemble learning framework for predicting student depression not only outperforms traditional models in accuracy, recall, and AUC but also demonstrates superior computational efficiency and stability, enhancing its engineering practicality. This method provides an expandable technical pathway and practical application potential for constructing intelligent mental health screening systems in higher education institutions.

5 CONCLUSION

This paper proposes a collaborative optimization approach for predicting student depression risk based on ensemble learning and feature engineering, establishing a systematic framework spanning data preprocessing, feature selection and dimensionality reduction, heterogeneous model fusion, and result output. Experimental results demonstrate that the model integrating LDA dimensionality reduction with multi-model blending outperforms traditional single models in metrics such as AUC, accuracy, and recall, validating its effectiveness and robustness in mental health prediction tasks.

Despite achieving favorable outcomes, limitations remain: first, the singular data source restricts sample distribution to specific institutions and populations, potentially affecting model generalization; second, features predominantly rely on static questionnaire data, lacking dynamic modeling support from time-series behavioral characteristics. Future research may enhance feature dimensional richness, model perceptual capabilities, interpretability, and real-time prediction capacity.

Overall, this study provides an algorithmic pathway for intelligent mental health screening among students that is both scalable and engineering-ready, laying a technical foundation for subsequent psychological risk early warning and personalized intervention.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] World Health Organization. Depression and Other Common Mental Disorders: Global Health Estimates. Geneva: WHO, 2022.
[2] Yang B X, Guo Y R, Hao S J, et al. Application of graph neural networks with data augmentation and ensemble learning strategies for depression detection. Computer Science, 2022, 49(07): 57-63.
[3] Teoh C-W, Ho S, Dollmat K S B, et al. Ensemble-learning techniques for predicting student performance on video-based learning. International Journal of Information and Education Technology, 2022, 12(8): 741-745.
[4] Jiang H, Hu R, Wang Y J, et al. Depression prediction in heart failure patients based on stacked models. World Journal of Clinical Cases, 2024, 12(21): 4661-4672.
[5] Vázquez-Romero A, Gallardo-Antolín A. Automatic detection of depression in speech using ensemble convolutional neural networks. Entropy, 2020, 22(6): 688.
[6] Feng W, Gou J, Fan Z, et al. An ensemble machine learning approach for classification tasks using feature generation. Connect Science, 2023, 35(1): 2231168.
[7] Pandey M, Taruna S. A comparative study of ensemble methods for students' performance modeling. International Journal of Computer Applications, 2014, 93(8): 1-6.
[8] Sun Y, Li Z, Li X, et al. Classifier selection and ensemble model for multi-class imbalance learning in education grants prediction. Applied Artificial Intelligence, 2021, 35(4): 290-303.

[9] Khan I, Gupta R. Early depression detection using ensemble machine learning framework. International Journal of Information Technology, 2024, 16: 3791-3798.
[10] B J, R J A K, Mitra A, et al. Education data analysis using ensemble models. In: Proceedings of the 4th International Conference on Smart Systems and Inventive Technology, 2022.
[11] Owen V E, Baker R S. Fueling prediction of player decisions: foundations of feature engineering for optimized behavior modeling in serious games. Technology, Knowledge and Learning, 2020, 25(2): 225-250.
[12] Janardhan N, Kumaresh N. Improving depression prediction accuracy using Fisher score-based feature selection and dynamic ensemble selection approach based on acoustic features of speech. Traitement du Signal, 2022, 39(1): 77-90.
[13] DESGM Authors. Enhancing depression detection: a stacked ensemble model with feature selection and RF feature importance analysis using NHANES data. Applied Sciences, 2024, 14(16): 7366.
[14] Hodge V J, Austin J. A survey of outlier detection methodologies. Artificial Intelligence Review, 2004, 22(2): 85-126.
[15] Sagi O, Rokach L. Ensemble learning: a survey. WIREs Data Mining and Knowledge Discovery, 2018, 8(4): e1249.
[16] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16: 321-357.
[17] Onan A. A stacked ensemble approach for text-based depression detection on social media. Expert Systems with Applications, 2022, 206: 117799.