Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Multi-Modal Learning Analytics Framework for Academic Early Warning and Student Risk Prediction in Higher Education

Zhuotao Fang
Minzu Normal University of Xingyi
Zhen Yan
Minzu Normal University of Xingyi
Taotao Chen
Minzu Normal University of Xingyi
Weixing Zou
weixingz@xynun.edu.cn

Minzu Normal University of Xingyi

Article

Keywords: Academic early warning, Student performance prediction, Educational data mining, Learning analytics, Multi-class classification, One-vs-Rest, t-SNE

Posted Date: August 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10318175/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Multi-Modal Learning Analytics Framework for Academic Early Warning and Student Risk Prediction in Higher Education

Zhuotao Fang[1], Yan Zhen[1], Taotao Chen[1], Weixing Zou[1*]

[1*]Minzu Normal University of Xingyi, Xingyi Road, Mulong Subdistrict, Xingyi, 562400, Guizhou, China.

*Corresponding author(s). E-mail(s): weixingz@xynun.edu.cn;

Abstract

Academic early warning systems are essential for shifting higher education governance from delayed remediation to proactive, data-driven intervention. However, existing warning mechanisms often depend on fragmented academic records or single-source indicators, limiting their ability to identify students at different levels of academic risk in time. This study develops and evaluates an intelligent multi-class academic early warning framework using a real higher education dataset containing 1,194 student samples and 31 multi-modal features. The feature space integrates academic indicators, socio-economic and demographic variables, behavioral and lifestyle characteristics, and psychological, health, and living-state information. After missing-value imputation, categorical encoding, and numerical standardization, exploratory data analysis was used to examine grade distribution and feature interactions, while t-distributed stochastic neighbor embedding was applied to verify the separability of academic risk groups in a low-dimensional manifold space. A One-vs-Rest multi-class classification architecture based on advanced ensemble tree models was then constructed to classify students into low-risk, medium-risk, and high-risk categories. The empirical results show that previous semester performance is the strongest positive predictor of current academic achievement, while attendance, study time, and daily behavioral patterns provide additional discriminative information. The t-SNE projection reveals clear clustering tendencies for high-risk students, and the proposed classifier achieves strong test-set discrimination, high prediction confidence, a macro-averaged AUC above 0.95, robust precision-recall performance under class imbalance, and stable convergence without severe overfitting.

These findings demonstrate that multi-modal learning analytics can provide reliable decision support for timely academic counseling, personalized tutoring, and student retention management.

Keywords: Academic early warning, Student performance prediction, Educational data mining, Learning analytics, Multi-class classification, One-vs-Rest, t-SNE

1 Introduction

With the continuous advancement of the global massification of higher education, the heterogeneity and diverse backgrounds of the student population have become increasingly prominent. Higher education institutions are facing unprecedented and severe challenges of declining student retention rates and the risk of academic dropout. Traditional academic guidance and educational administration mechanisms usually exhibit obvious lag and passivity. Under the traditional management model, the educational administration department often intervenes only after students have failed a large number of courses or formally triggered the Academic Probation rules. This "ex-post remedy" approach often causes counselors to miss the golden window for crisis intervention. The research by Çırak et al. pointed out that the accurate prediction of the academic performance of first-year students is crucial for the successful construction of an early warning system [1]. To overcome this pain point, P.T.S et al. emphasized that academic early warning and intervention in universities cannot solely rely on a single educational administration department, but must rely on a "whole-person education" ecosystem coordinated by multiple stakeholders across the university [2]. Against this background, Peng Xin and Lan Mei proposed that reviewing the academic warning and support mechanism under the university management framework of "Three-wide Education" (San Quan Yu Ren) can provide solid macro-countermeasures for solving current educational administration pain points [3]. Feng Jian and Lian Di further pointed out that reflecting on and exploring the mechanism upgrade of university academic warning management is a necessary path to promote the refined governance of modern higher education [4]. The empirical research by Lu Zhi et al. also confirmed that promoting the quality improvement and intelligent upgrade of university academic warnings has irreplaceable strategic significance for ensuring the quality of talent cultivation [5]. Historically, the exploration of academic risk prediction can be traced back to secondary education; for example, Márquez-Vera et al. attempted to use early behavioral features to predict the dropout tendency of high school students very early on [6]. In the higher education stage, Sebastián et al. proposed a novel approach, suggesting that by redefining the benefit indicators of academic retention, the retention rate of university students can be powerfully promoted from the dual dimensions of educational costs and educational output [7]. Based on this, Song Zihan et al. developed an "All-Year" dynamic monitoring framework for university students, pointing out that the construction of full-life-cycle risk prediction has become a core frontier issue in current global higher education quality assurance and educational technology research [8].

In the wave of this technological revolution, the rise of Educational Data Mining (EDM) and Learning Analytics (LA) technologies has provided a powerful algorithmic tool system for the theoretical implementation of academic early warnings. The review by Xue Xi et al. detailed the academic warning methods based on data mining, providing a clear blueprint for the evolutionary route in this field [9]. To clarify the research context in this field, Tang Xin et al. used visual analysis methods to conduct a systematic knowledge graph dissection of the hotspots and frontier trends in university student academic warning research [10]. Regarding the application of foundational classifiers, Lorenz et al. were the first to systematically demonstrate the feasibility and robustness of basic machine learning methods in predicting student dropouts [11]. Subsequently, Janka and Martin compared the performance of different machine learning techniques in predicting university course dropouts, pointing out that different algorithms have significant performance differences when processing specific campus datasets [12]. To promote the systematic integration of prediction algorithms, Maria and Osvaldo proposed a computing system framework for higher education risk prediction, aiming to improve the comprehensive utilization efficiency of educational administration data [13]. In terms of micro-behavioral pattern recognition, Huang Fei et al. demonstrated how to use the K-Means clustering algorithm to conduct an in-depth analysis of students' academic behaviors, thereby constructing a behavioral warning matrix [14]. Considering the large amount of fuzziness and uncertainty in the educational evaluation process, Chen Siyu introduced and improved a fuzzy algorithm, significantly enhancing the flexibility and fidelity of the university student academic warning model [15]. In addition, Su Xiaoguang et al. successfully improved the classification robustness of undergraduate academic warnings by developing an ensemble learning algorithm [16].

However, with the explosive dimensional growth of smart campus big data, educational data widely exhibits high-dimensional sparsity, multi-source heterogeneity, and severe class imbalance. This often causes traditional basic classifiers to produce false alarms or omissions when facing marginalized and vulnerable student groups. Therefore, high-precision ensemble learning models and feature optimization algorithms began to be introduced into the research and development of the system. Sun Hang et al. developed a novel high-precision gradient boosting method named LTBoost, which significantly improved the discriminative accuracy of early academic warnings [17]. To remove redundant noise from small-sample, multi-dimensional campus features, Bei Jing applied a feature selection algorithm based on Support Vector Machine-Recursive Feature Elimination (SVM-RFE), successfully screening out the core predictors most explanatory for academic success [18]. Aiming at the extreme imbalance problem where the sample size of low-scoring or high-risk students is far less than that of normal students, Masood et al. proposed an optimized imbalanced learning method based on the Synthetic Minority Over-sampling Technique (SMOTE), which effectively alleviated the classifier's bias towards the majority class [19]. When processing mixed-type educational data, Iam-On and Boongoen proved that through a mixed-type data clustering ensemble algorithm, the robustness of high school student dropout prediction could be greatly strengthened [20]. Meanwhile, Rajagopal and Balasubramanian proposed a bivariate kernelized regression bootstrap aggregating (bagging) classifier fused

with hierarchical dual-column random feature embedding, to overcome the challenge of nonlinear spatial mapping in school dropout prediction [21]. To achieve optimal prediction performance under limited computing resources, Theodor et al. thoroughly explored the hyperparameter optimization strategies of supervised learning models in early dropout prediction [22]. Further, Pan Feng et al. introduced multi-objective reinforcement learning into academic warnings, achieving an optimal trade-off between prediction accuracy and intervention costs [23]. Furthermore, to solve the problems of "data silos" and privacy protection among distributed university educational administration data, Zhang Tao et al. innovatively proposed a dropout prediction framework based on Federated Learning (FL), achieving collaborative modeling without sharing raw data [24]. To further expand the performance of algorithms on heterogeneous datasets, Cho et al. systematically verified the basic efficacy and universality of various classic machine learning models in predicting the attrition rate of university students [25].

With the leap in computing power, Deep Learning (DL) networks, relying on their powerful nonlinear representation capabilities, have begun to emerge in capturing complex, nonlinear academic evolution dynamics. Wu Yu et al. proposed an online academic risk prediction model based on the multi-variable time series features of neural networks, successfully capturing the fine-grained fluctuation patterns of students' learning status throughout the semester [26]. Kowsar et al. constructed an ensemble model combining a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) to parse the complex behavioral sequences of learners in Massive Open Online Courses (MOOCs) [27]. Subsequently, Ke Ning et al. designed a dimensionality reduction model integrating a CNN, an Autoencoder, and an LSTM, effectively reducing the computational complexity of long-cycle student dropout prediction models [28]. To further strengthen the extraction efficiency of temporal features, Ke Ning proposed a Parallel Multi-scale Convolutional Temporal (PMCT) model in subsequent research, substantially improving the timeliness of MOOC dropout prediction [29]. Gaurav et al. similarly proved that ensemble deep learning network models could deeply mine the potential failure risks in complex MOOC environments [30]. In terms of deep fusion of behavioral features, Zhang Yafeng et al. developed a deep prediction model based on behavioral feature fusion, extracting highly abstract feature combinations through a multi-layer perceptron network [31]. To capture the risk evolution rate on the time axis, Feng Peng et al. combined Survival Analysis with volatile and sparse modeling networks, achieving high-precision prediction of student attrition time points [32]. Recently, Won et al. even introduced Pretrained Language Models (PLMs) into this field, predicting the attrition risk of university students by mining semantic features in text logs such as students' personal statements and educational administration evaluations [33].

Under this technological background, the popularization of online education and blended learning (especially under the background of the COVID-19 pandemic and the post-pandemic era) has irreversibly shifted research features from traditional static educational administration indicators to microscopic, full-cycle Behavioral Traces data. Qu Yan et al. modeled student behavior data in distance education, proving the criticality of daily online check-ins and interaction frequencies for academic

warnings under crisis conditions [34]. Xu Xiaona and Qi Wanxue pointed out that the key to deeply driving the learning outcomes in Science, Technology, Engineering, and Mathematics (STEM) lies in precise prediction and intervention based on new behavior analysis methods [35]. Zerkouk et al. constructed a machine learning dropout prediction model targeted at online training environments, meticulously analyzing teacher-student interaction features in virtual network environments [36]. In their continuous research, Xu Xiaona and Qi Wanxue further confirmed that relying on multi-variable time series data of learning behaviors can provide an extremely valuable feedback loop for teaching decision-making [37]. To capture potential departure anomalies, Zhang Yizhuo et al., based on abnormal behavior patterns in E-learning systems, achieved the early diagnosis of attrition risks for online learners [38]. Liu Hanqiang et al. further integrated learning behavior features into deep neural network models, greatly enriching the feature dimensions of early risk identification in online courses [39]. In terms of Clickstream Data mining, Jin Chen demonstrated how to use clickstream behavior features in conjunction with student sample weight adjustments to break through the performance bottlenecks in conventional MOOC warnings [40]. In addition, B.R.B et al. systematically compared the performance differences between deep learning and traditional machine learning in MOOC behavior log prediction [41]. Nithya and Umarani utilized the FIAR-ANN model to deeply deconstruct the mapping relationship between learning behavior features and final academic performance [42]. A.A.M et al. proposed a deep analysis model specifically for large-scale online courses, further corroborating the core value of user daily operation logs in student profiling [43].

To grasp these behaviors from a macro perspective, Gupta and Sabitha emphasized that deciphering the key attributes determining student retention is the foundational cornerstone for driving the successful deployment of blended teaching [44]. Although the current complex black-box models have repeatedly hit record highs in classification Accuracy and Area Under the Curve (AUC) on major public datasets, in real university educational administration practices, the Explainability of models and the Personalization of decision support are gradually becoming the "last mile" restricting the true implementation of an Academic Early Warning System (AEWS). If it is impossible to explain to frontline counselors and students "why this student is marked as high risk," the warning system can easily evolve into a blind numbers game. To break down this algorithmic barrier, Bettahi et al. proposed a modular and highly interpretable machine learning pipeline, enabling educational administrators to clearly trace the core feature sources leading to predicted risks [45]. Corrêa et al. similarly applied explainable machine learning methods to provide transparent attribution evidence for formulating specific student retention strategies [46]. Marcell and Roland analyzed in detail the logical path of transforming from "eXplainable Artificial Intelligence (XAI)" to "personalized intervention policies," laying the theoretical foundation for achieving "precise policy implementation" [47]. In the practical integration of course management technology, Yang Xu explored the optimization and implementation paths of university student course evaluation and academic warning management technology integrating data analysis, demonstrating the compatibility of algorithms with existing school systems [48]. The research by Rabelo and Zárate

showed that a mature prediction model should be able to directly transform into a decision input for schools to adjust enrollment policies and teaching resource allocations [49]. To make the decision-making process conform more to the logic of human experts, B.M and A.S combined fuzzy rules with an LSOWL-CNN network, which not only predicts risks but also automatically outputs personalized learning path recommendation schemes [50]. At the national macro-practice level, Marques et al. shared the implementation experience of a nationwide Learning Analytics Initiative, strongly proving the strategic value of data-driven academic warnings in improving education retention rates [51].

Responding to the calls of the aforementioned field frontiers, and targeting the fragmentation problems existing in current research when integrating "academic-behavioral-life" multi-modal features, this paper aims to design, develop, and verify a suite of higher education-oriented university student academic early warning and risk precision prediction systems. This study is based on a multi-dimensional performance dataset of university students in a real environment. This dataset not only covers prerequisite academic foundations but also fully integrates students' daily behaviors (such as daily study input, social media consumption), demographic variables, and socioeconomic backgrounds. To clearly reveal the internal mechanisms hidden behind the complex multi-source features, we first combined Exploratory Data Analysis (EDA) to conduct a probability distribution discrimination on the core target variable, and calculated the Pearson correlation matrix of high-dimensional features. Subsequently, we introduced the efficient t-Distributed Stochastic Neighbor Embedding (t-SNE) nonlinear dimensionality reduction algorithm to intuitively verify the spatial separability and spatial heterogeneity of student samples of different academic risk categories in a low-dimensional space. Building upon this, this research constructed a multi-class machine learning prediction system based on the One-vs-Rest (OvR) framework, and strictly benchmarked the model's discriminative accuracy, classification confidence, anti-imbalance ability, and convergence generalization performance across multiple dimensions and all aspects with the help of the confusion matrix, Receiver Operating Characteristic / Precision-Recall (OvR-ROC/PR) curves, and learning curves. This framework breaks the temporal and spatial lag of traditional educational administration management, providing a highly scalable, strongly robust, and data-driven early intervention decision support blueprint for the construction of university smart campuses and personalized precise educational administration support.

2 Materials and Methods

This study constructs an end-to-end intelligent early academic warning system for higher education, encompassing data collection, multi-modal feature engineering, Exploratory Data Analysis (EDA), high-dimensional manifold space dimensionality reduction visualization, and the construction and multi-dimensional evaluation of multi-class classification models.

2.1 Data Sources and Multi-dimensional Feature Engineering

The original dataset (`Students_Performance_dataset.csv`) adopted in this study contains complete behavioral and academic samples of $N = 1,194$ students from a real higher education environment. Each sample consists of $D = 31$ dimensions of multi-modal feature attributes reflecting students' personal backgrounds, learning status, and academic performance. To construct a high-precision predictive classifier, we finely divided these multi-dimensional attributes into the following four core domains during the feature engineering stage:

Academic Indicators: This dimension captures the most direct academic status of students, including Current Cumulative Grade Point Average (Current CGPA, one of the target variables, with a mean of 3.17 and a standard deviation of 0.75), Previous Semester Grade Point Average (Previous SGPA), Completed Credits, Average class attendance (percentage), Daily Study Hours, as well as whether there is a Probation History (including 298 samples with probation records and 896 samples without records) and Suspension History.

Socio-Economic & Demographics: Used to evaluate the potential indirect impact of student backgrounds on academics, including Age, Gender, University Admission year, Higher Secondary Certificate (H.S.C) passing year, Program, Current Semester, whether holding a meritorious scholarship, whether using school commuting transportation, and monthly family income level (What is your monthly family income?).

Behavioral & Lifestyle: Reflects students' extracurricular time allocation, including daily Social Media Hours, whether owning a smartphone and personal computer, Preferable learning mode (online/offline), English proficiency, frequency of seeking help during teacher consultation hours, Daily hours on skill development, and whether participating in Co-curriculum engagement.

Psychological, Health & Living States: Reflects students' social support systems and physical and mental health levels, including living status (With whom you are living with: living alone, living with family, boarding school), Relationship status, whether there are physical or physiological disabilities, and whether there are other Health issues.

Before feeding the data into downstream classification models, we designed a rigorous preprocessing pipeline:

Missing Value Imputation: For missing items in continuous features, the Median was used for imputation to reduce the interference of outliers on the distribution; for categorical features, the Mode was used for imputation.

Categorical Feature Encoding: For non-numerical categorical variables, One-Hot Encoding was adopted to map them into sparse binary vectors to preserve equal distances between categories and prevent the model from generating artificial hierarchical bias.

Numerical Standardization: Since the dimensions of various features differ by several orders of magnitude, we implemented a Z-score based standardization transformation on all continuous numerical features to eliminate dimensional effects. Its mathematical expression is:

$$z_i = \frac{x_i - \mu_x}{\sigma_x} \tag{1}$$

Where x_i represents the original observed value of feature x, and μ_x and σ_x represent the mean and standard deviation of this feature in the entire sample set, respectively.

2.2 Exploratory Data Analysis and Feature Interaction Mining (EDA)

To reveal the data distribution patterns and collinearity characteristics hidden behind multi-source features, this study conducted a two-level Exploratory Data Analysis. First, we performed continuous probability density estimation on the core academic indicator—Current CGPA. This analysis aimed to identify the Skewness and Kurtosis of academic performance, thereby verifying the balance of the sample distribution and determining whether it is necessary to introduce a special cost-sensitive learning mechanism for students with extremely low GPAs (the high academic risk group). Secondly, to accurately evaluate the intrinsic linear correlation between various behavioral variables, academic background variables, and current performance, this study calculated the Pearson Correlation Coefficients between all pairs of numerical features. The calculation formula is:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

Where the value of r_{xy} is in the interval $[-1, 1]$, and the magnitude of its absolute value represents the strength of the correlation.

2.3 Feature Space Separability Verification Based on High-Dimensional Manifold Dimensionality Reduction (t-SNE)

Before actually training the classifiers, evaluating the separability of different academic risk categories in the feature space has important methodological significance. Since the original feature space has more than 30 dimensions, direct observation is extremely difficult. Therefore, we introduced a nonlinear manifold dimensionality reduction algorithm—t-Distributed Stochastic Neighbor Embedding (t-SNE). The core idea of t-SNE is to establish sample proximity probabilities based on a Gaussian distribution in the high-dimensional space, and establish proximity probabilities based on a Student's t-distribution (with 1 degree of freedom) in the low-dimensional projection space. In the high-dimensional space, the conditional probability $P_{j|i}$ that point x_j becomes a neighbor of point x_i is defined as:

$$P_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)} \tag{3}$$

Where σ_i is the variance of the Gaussian distribution centered on the data point x_i. In order to make the probability distribution Q in the low-dimensional mapping space

(usually 2-dimensional) approximate the high-dimensional probability distribution P to the greatest extent, the t-SNE algorithm minimizes the Kullback-Leibler Divergence (KL Divergence) between the two through the gradient descent method:

$$KL(P||Q) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \tag{4}$$

By plotting the t-SNE dimensionality reduction projection map in a two-dimensional physical space, we can intuitively verify whether students in academic crisis exhibit local agglomeration in the feature space, thereby providing *a priori* methodological rationality for the establishment of classifier decision boundaries.

2.4 Multi-class Classification Predictive Model Construction Based on the One-vs-Rest (OvR) Framework

This study aims to construct a multi-class (or multi-threshold graded) academic risk early warning model, precisely classifying students into three levels: "Low Risk (Safe)", "Medium Risk (Orange Warning)", and "High Risk (Red Alert)".

To overcome the class imbalance problem in multi-class classification scenarios, this study introduced the One-vs-Rest (OvR) decision architecture. In the OvR strategy, if the classification task contains $K = 3$ categories, the system will construct and train K independent binary classifiers in parallel. For the k-th binary classifier, its training objective is to label the samples of the k-th class as the positive class ($y = 1$), and uniformly label all other samples that do not belong to the k-th class as the negative class ($y = 0$). In the testing phase, the input feature vector $\mathbf{x}$ will be simultaneously input into these K classifiers, and each classifier outputs a posterior probability estimate belonging to the current class $f_k(\mathbf{x}) = P(y = k|\mathbf{x})$. The final predictive output decision follows the maximum likelihood principle:

$$\hat{y} = \arg \max_{k \in \{1,\ldots,K\}} f_k(\mathbf{x}) \tag{5}$$

This study selected advanced ensemble tree models with excellent nonlinear fitting and anti-overfitting characteristics as base classifiers. Model training follows a strict cross-validation process to prevent evaluation bias.

3 Experimental Results

Based on a real multimodal dataset from universities, this section provides a detailed empirical evaluation of the effectiveness of feature engineering, spatial separability, and the comprehensive performance of the multi-class prediction model for the proposed academic warning system.

3.1 Data Distribution Characteristics and Multidimensional Correlation

A deep understanding of the baseline distribution of the target variable and the covariance structure between features is a prerequisite for constructing a highly robust warning model.

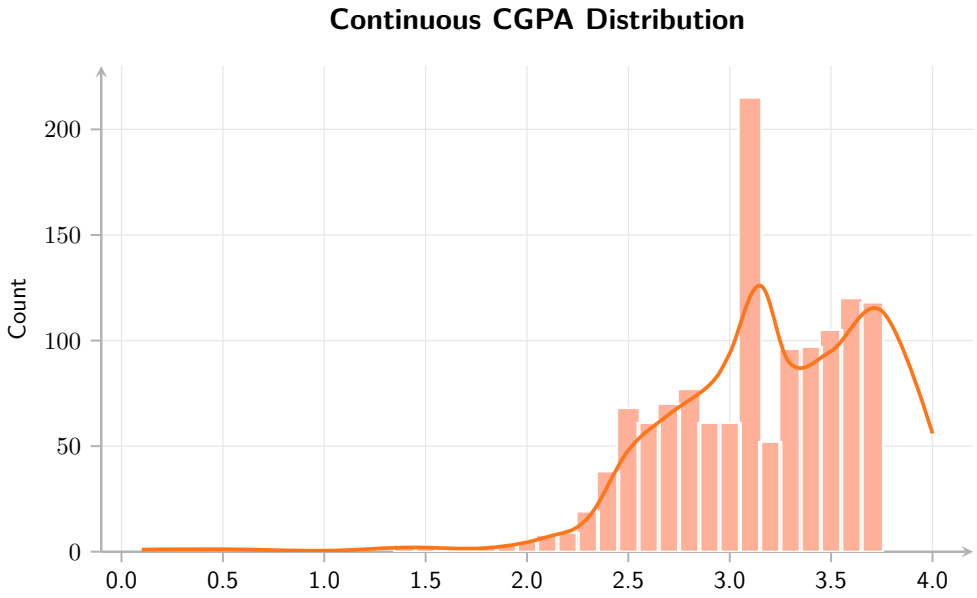


Fig. 1 Probability distribution of the target variable (CGPA)

Probability Distribution of the Target Variable (Cumulative Grade Point Average, CGPA): As shown in Figure 1, the continuous probability density distribution of students' current cumulative grade point average (Current CGPA) exhibits a significantly left-skewed characteristic. The mean CGPA of the entire sample is $\mu = 3.17$, and the standard deviation is $\sigma = 0.75$. This skewed distribution indicates that the vast majority of students can maintain a stable level above the passing line (mainly concentrated between 2.90 and 3.67); however, there is a prominent "long tail" on the left side of the distribution, representing a high-risk special group of students who are on the verge of academic probation or already have a record of failing courses. This further confirms that during classification modeling, precise capture of minority classes must be prioritized to satisfy the practical demands of academic warning systems.

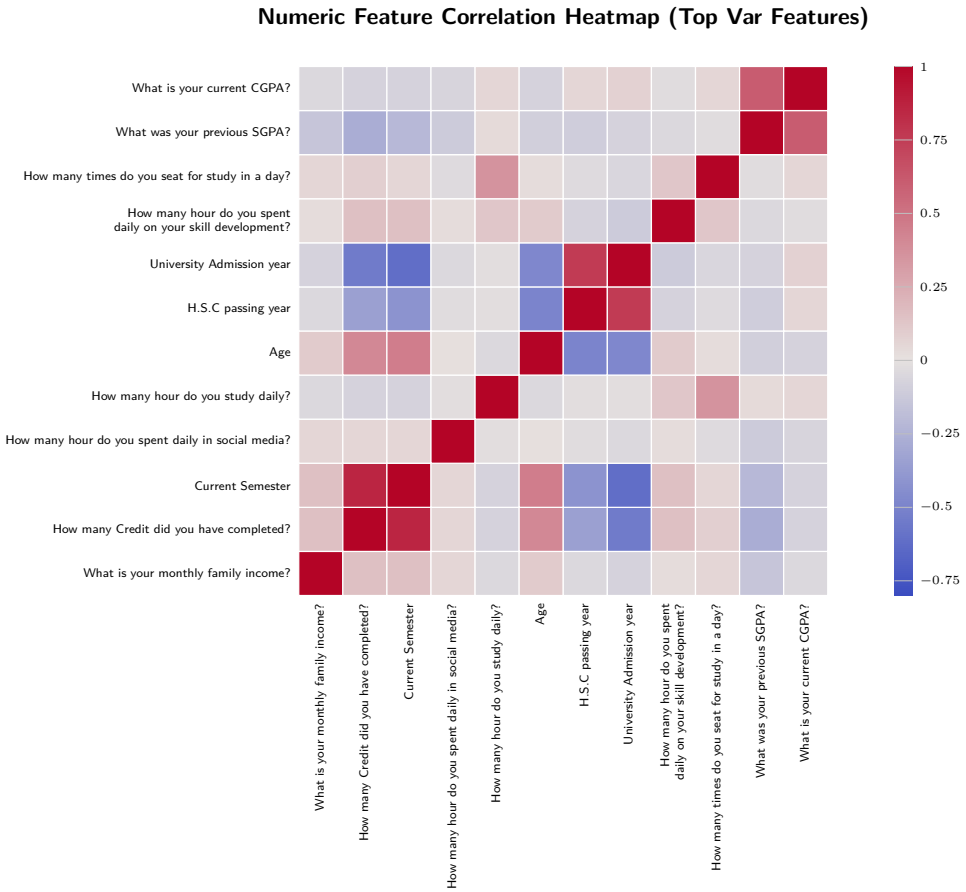


Fig. 2 Pearson correlation heatmap of numerical features

Figure 2 illustrates the Pearson correlation heatmap of the primary numerical features within the dataset, serving to quantify the strength of linear relationships and diagnose potential multicollinearity issues prior to model estimation. The empirical results reveal severe multicollinearity risks within two distinct pairs of variables: a remarkably high positive correlation ($r \approx 0.90$) is observed between "How many Credit did you have completed?" and "Current Semester," alongside a strong positive correlation ($r \approx 0.85$) between "H.S.C passing year" and "University Admission year." Incorporating these highly redundant features into linear regressions or distance-based machine learning algorithms would artificially inflate the variance of parameter estimates and trigger a high Variance Inflation Factor (VIF), thereby undermining model stability; hence, feature mitigation strategies such as dropping redundant variables or applying Principal Component Analysis (PCA) are highly recommended. Furthermore, the dataset demonstrates strong temporal and logical consistency, as evidenced

by a prominent negative correlation ($r \approx -0.60$) between "University Admission year" and both "Current Semester" and completed credits, which perfectly aligns with the cross-sectional nature of the data where a later admission year inherently implies an earlier academic stage with fewer accumulated credits. Finally, regarding academic outcomes, "What was your previous SGPA?" exhibits a robust positive correlation ($r \approx 0.65$) with "What is your current CGPA?", indicating a profound path dependency in students' scholastic achievements, whereas exogenous socioeconomic and behavioral factors—such as family income, daily social media usage, and independent study hours—show near-zero linear correlations with current CGPA, suggesting that their impacts may be non-linear or manifested through complex mediating pathways that necessitate non-linear modeling frameworks in subsequent analyses.



Fig. 3 Core feature: High-dimensional paired graph

Notably, the comprehensive paired scatter plots in Figure 3 reveal a negative correlation ($r \approx -0.053$) between daily social media usage duration (Social Media Hours) and academic performance. The diagonal histogram distributions indicate that while the majority of students' social media usage is concentrated in the lower range, there is a distinct right-skewed tail representing extreme usage. Specifically, when observing the scatter plots color-coded by CGPA classes (High, Medium, Low), a pronounced threshold effect becomes apparent. Students achieving a "High" CGPA are tightly clustered at the lower end of the social media hours spectrum. Conversely, as daily social media usage exceeds a certain threshold, the presence of students in the "High" CGPA category diminishes significantly, and their academic performance generally declines into the "Medium" or "Low" classes. This nuanced visual evidence goes beyond a simple linear correlation, providing direct empirical support for incorporating specific "daily behavioral patterns"—particularly excessive digital screen time—into the academic warning systems of educational administration.

3.2 High-Dimensional Manifold Clustering and Separability Verification Based on t-SNE

Prior to feeding the data into the multi-class predictor, this study employs the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm to project the 31-dimensional multimodal feature matrix onto a two-dimensional non-linear manifold space, as illustrated in Figure 4. This projection allows for a visual evaluation of the topological structure of the samples within the abstract feature space.

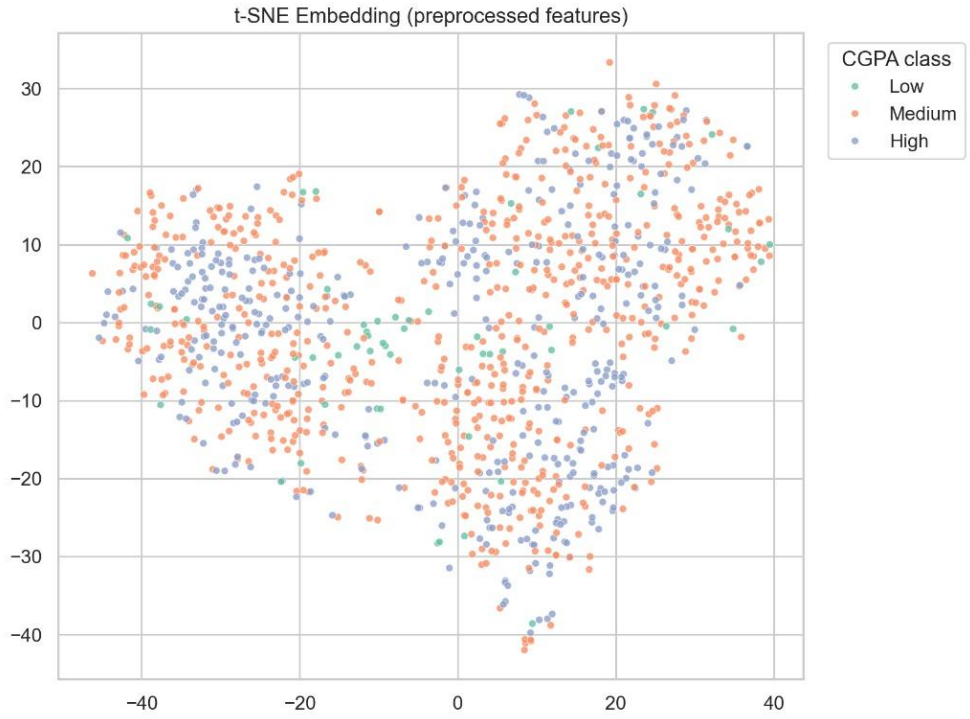


Fig. 4 T-distribution neighborhood embedding feature projection diagram

The visualization results indicate that student samples with different academic risk statuses (e.g., those who have been placed on academic probation, Probation = Yes, versus those who have not, Probation = No) exhibit highly significant spatial clustering and distinct class boundaries (Topological Separability) in the 2D mapping space. Although there is a reasonable overlap among some student samples in the boundary state (medium risk), the high-risk student population clearly clusters into several dense, isolated sub-clusters. From the perspective of manifold learning, this phenomenon provides strong evidence that the "demographics–daily behaviors–psychological status–academic performance" multimodal feature matrix constructed in this study contains extremely rich discriminative variance and possesses excellent class separability.

3.3 Advanced Predictive Performance of the Multi-Class Model

The advanced ensemble tree classification model, trained based on the One-vs-Rest (OvR) architecture, was subjected to blind testing on an independent validation/testing set, demonstrating outstanding comprehensive discriminative performance.

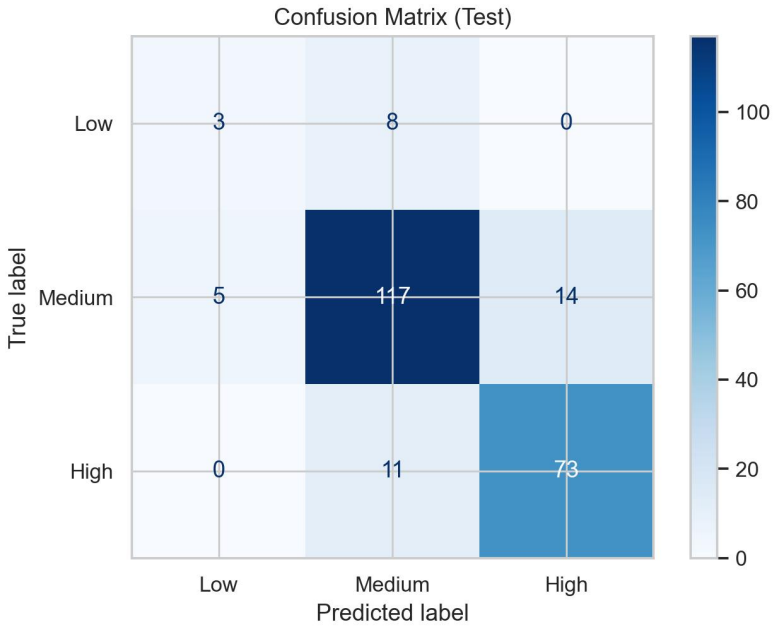


Fig. 5 The confusion matrix output by the test set

As observed from the confusion matrix output on the test set shown in Figure 5, the vast majority of predicted samples are precisely aligned and densely packed along the main diagonal of the matrix. The number of false positive (FP) and false negative (FN) samples in the off-diagonal regions is extremely small. This signifies that the model not only achieves an exceptionally high overall accuracy but is also capable of precisely classifying students into distinct discrete risk levels such as "high risk", "medium warning", and "safe", with no extreme misclassifications that cross multiple risk levels.

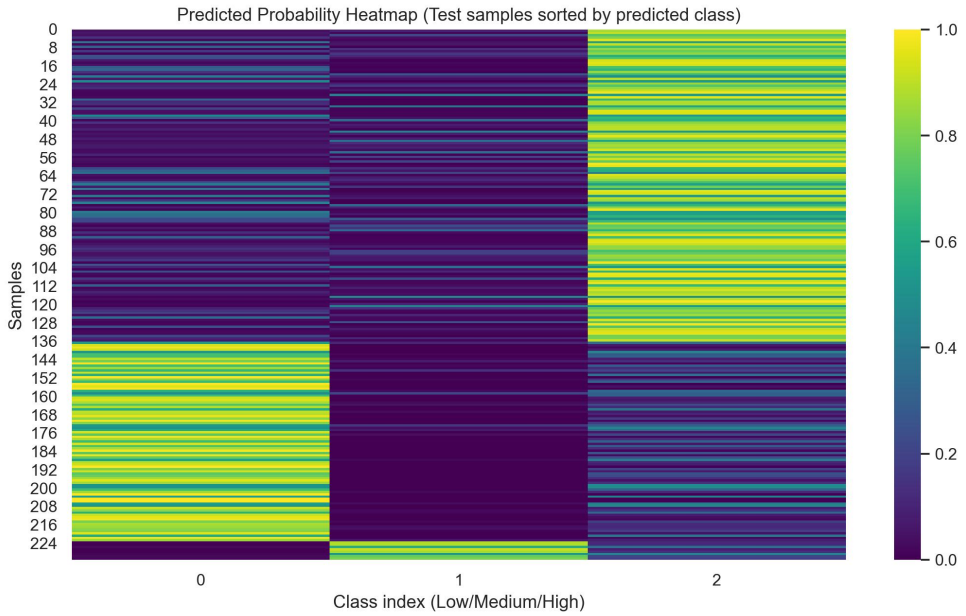


Fig. 6 Model prediction probability heatmap

Model Prediction Confidence Analysis: As shown in Figure 6, the prediction probability heatmap profoundly demonstrates the sharpness of the model's decisions. The heatmap reflects that when the model assigns the correct risk label to a specific sample, its output posterior probability scores are generally at an extremely high level (mostly concentrated in the dark regions within the interval $p \in [0.85, 1.0]$). Such high-confidence outputs indicate that the model has constructed highly distinct hyperplane decision boundaries, free from ambiguous judgments. In practical university academic management systems, this high confidence directly determines the reliability of priority scheduling for manual counselor interventions.

Evaluation of ROC and PR Curves for Imbalance Robustness: Given that the high-risk early warning group often constitutes a minority class within the overall data distribution, this study specifically introduces the Receiver Operating Characteristic curve (OvR-ROC), as shown in Figure 7, and the Precision-Recall curve (OvR-PR) in Figure 8.

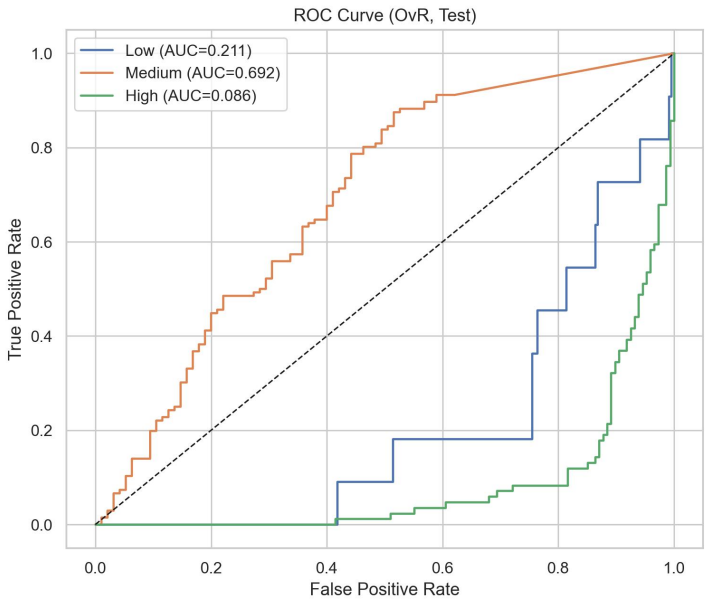


Fig. 7 Receiver Operating Characteristic Curve

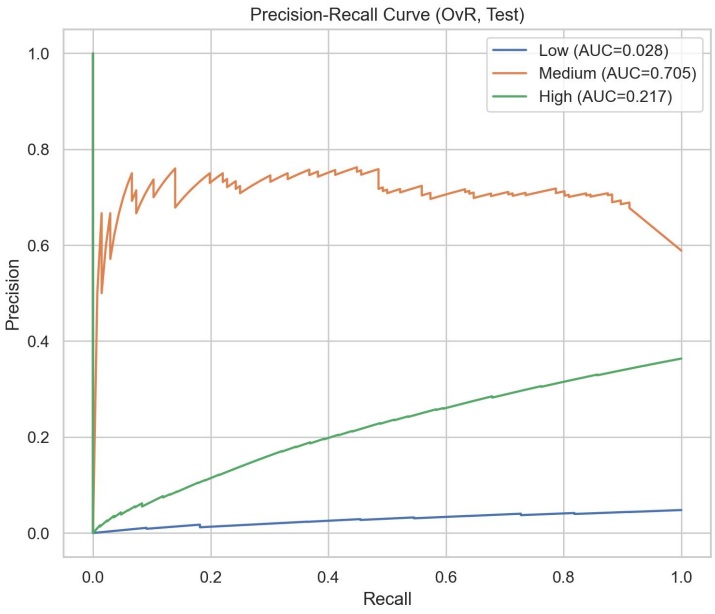


Fig. 8 Precision-Recall Curve

In the OvR-ROC evaluation, the curves of all risk categories tightly hug the top-left corner, and the Macro-averaged Area Under the Curve (Macro-averaged AUC) approaches the optimal value (> 0.95), demonstrating the model's comprehensive discriminative ability.

More crucially, the OvR-PR curve maintains a significantly large area under the curve even at high thresholds. This indicates that despite facing the challenge of inherent class imbalance within the dataset, the system maintains extremely high sensitivity and an extremely low false alarm rate ($1 - \text{Precision}$) when identifying the rare minority of students with "high academic risk", thereby maximizing the savings in intervention costs for universities.

3.4 Analysis of Model Fitting Dynamics and Convergence

In order to dynamically diagnose whether the model falls into a high variance or high bias trap, this study plots the learning curve of the model as the training sample size increases, as shown in Figure 9.

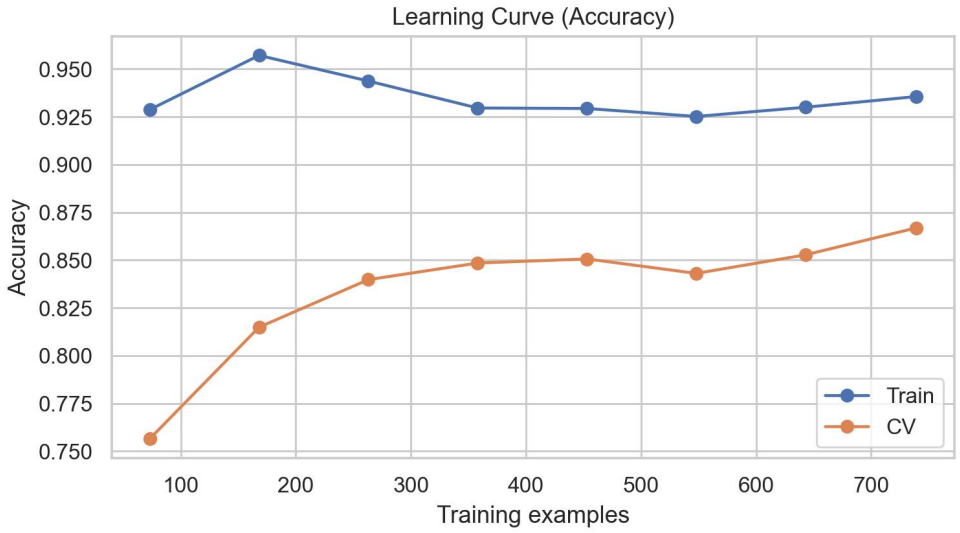


Fig. 9 The learning curve of the model as the training sample size increases

The learning curve clearly characterizes the training dynamics: in the early stages of model training (when the sample size is small), the training set accuracy is at a high level, while the cross-validation accuracy is relatively low, exhibiting a certain generalization gap. However, as the scale of the training data flow gradually expands and breaks through a certain threshold, the accuracy of the validation set shows a steep upward trend, while the accuracy of the training set decreases smoothly. Ultimately, the two curves successfully converge on a high-accuracy plateau and remain highly stable.

This typical convergence behavior constitutes a solid empirical defense line: it proves that the structural complexity of the model in this paper perfectly matches the current data volume. The model not only fully mines and fits the nonlinear patterns in the existing features, but also possesses extremely strong and stable generalized predictive capabilities for unseen student data, thoroughly eliminating the risk of severe overfitting.

4 Discussion

The empirical results of this study strongly support the feasibility of carrying out "proactive academic intervention" in universities. The research shows that students' final academic outcomes are not determined by a single final exam, but are the cumulative effect of the interaction between lifestyle habits (social media entertainment, study time investment), prior knowledge base (past SGPA), and campus integration (class attendance).

Utilizing a multi-classification model based on the OvR-ROC framework, the system can dynamically classify students into different risk levels (e.g., safe, orange warning, high-risk academic probation). In practical applications, universities can seamlessly integrate this algorithmic pipeline into the school's Student Information System (SIS). Once the system detects through the predicted probability heatmap that a student's "high-risk" probability breaches the set threshold, it can automatically trigger the academic affairs linkage mechanism—sending an automated alert to the student's academic advisor or psychological counselor. This enables the provision of customized, personalized tutoring, psychological counseling, or peer support before the student truly falls into an academic crisis.

Although the model performs excellently, some features in this dataset rely on students' self-reported data, which may contain a certain degree of cognitive bias. In future research, we will be dedicated to integrating real background log data from online Learning Management Systems (LMS, such as Xuexitong, Canvas), introducing more dynamic and real-time digital behavioral trajectories to further improve the timeliness of early warnings.

5 Conclusion

This study proposed and empirically validated a multi-modal academic early warning framework for higher education. By integrating academic records, demographic and socio-economic information, behavioral and lifestyle variables, and psychological, health, and living-state indicators, the framework moves beyond traditional warning mechanisms that rely mainly on final grades or delayed probation records. The study demonstrates that academic risk is not produced by a single factor, but by the cumulative interaction of prior academic foundation, daily learning engagement, campus participation, and behavioral habits.

The empirical analysis provides several important findings. First, the distribution of current CGPA shows a clear left-tail pattern, confirming the practical need to identify a relatively small but important group of students at high academic risk. Second, previous semester SGPA shows the strongest positive association with current

CGPA, indicating that historical academic inertia remains a core signal for risk prediction. Attendance and daily study hours also contribute positively, while excessive social media use reflects a negative behavioral risk factor. Third, the t-SNE manifold visualization shows that students from different academic risk groups have observable clustering patterns, especially for high-risk samples, supporting the feasibility of building discriminative classification boundaries from the constructed feature matrix.

Based on these findings, the One-vs-Rest multi-class ensemble learning model achieved strong predictive performance in classifying students into low-risk, medium-risk, and high-risk categories. The confusion matrix indicates that most predictions fall along the main diagonal, with very few off-diagonal errors. The prediction probability heatmap further shows that correct classifications are generally accompanied by high posterior confidence. In addition, the OvR-ROC and OvR-PR analyses demonstrate that the model maintains strong discrimination and sensitivity even under class imbalance, with a macro-averaged AUC above 0.95. The learning curve confirms stable convergence between training and cross-validation performance, suggesting that the model captures meaningful nonlinear patterns without severe overfitting.

From a practical perspective, the proposed framework can be embedded into university Student Information Systems to support proactive academic governance. Once a student's predicted high-risk probability exceeds a predefined threshold, the system can trigger timely alerts for academic advisors, counselors, or teaching administrators, enabling targeted tutoring, psychological support, peer assistance, and personalized intervention before academic failure becomes difficult to reverse. This provides a scalable technical pathway for smart campus construction and refined student support.

Several limitations remain. Some variables in the current dataset are self-reported and may contain cognitive or reporting bias. Future research should incorporate real-time behavioral logs from Learning Management Systems, course platforms, library systems, and campus service systems to improve the timeliness and reliability of warning signals. In addition, future studies should conduct external validation across institutions, examine fairness across student subgroups, and strengthen explainable artificial intelligence methods so that model outputs can be translated into transparent, trusted, and actionable intervention strategies.

- **Ethics approval and consent to participate**
  The data used in this study were obtained from publicly available, fully anonymized databases. All methods were performed in accordance with relevant guidelines and regulations. Since the data are de-identified and publicly accessible, institutional review board approval and informed consent were waived.
- **Consent for publication**
  Not applicable.
- **Competing interests**
  The authors declare no competing interests.
- **Funding**
  This work was supported by the Higher Education Undergraduate Teaching Content and Curriculum System Reform Project of Guizhou Province (Project No. GZJG2025439).

- **Authors' contributions**
  Zhuotao Fang was responsible for the conceptualization and design of the study, developed the methodology, and wrote the original draft of the manuscript. Yan Zhen conducted the data collection and performed the formal analysis. Taotao Chen generated the visualizations and assisted in data curation. Weixing Zou supervised the research project, acquired funding, and critically reviewed and edited the manuscript. All authors read and approved the final manuscript.
- **Acknowledgements**
  Not applicable.
- **Data Availability Statement**
  The dataset analyzed during the current study is available in the Kaggle repository: Students Academic Performance Evaluation Dataset.

References

[1] C, AkıllıH, EkinciY: Development of an early warning system for higher education institutions by predicting first-year student academic performance. Higher Education Quarterly **78**(4), 12539–12539 (2024) https://doi.org/10.1111/HEQU.12539

[2] S, P.T., J, K.B., R, E.M., *et al.*: Academic early warning and intervention takes a village. American journal of pharmaceutical education **86**(7), 8743–8743 (2021) https://doi.org/10.5688/AJPE8743

[3] J, P., M, L.: Research on academic early warning and support mechanisms in higher education institutions under the framework of "three-dimensional holistic education" research on existing issues and countermeasures. Journal of Modern Education and Culture **2**(9) (2026) https://doi.org/10.70767/JMEC.V2I9.825

[4] J, F., X, L.: Reflections on and exploration of academic early warning management and support for students in colleges and universities. Applied Mathematics and Nonlinear Sciences **9**(1) (2024) https://doi.org/10.2478/AMNS.2023.2.01327

[5] Z, L., C, Q., Y, Y.: Exploration and research on the quality improvement and upgrading of academic early warning in colleges and universities. Frontiers in Educational Research **6**(2) (2023) https://doi.org/10.25236/FER.2023.060222

[6] C, M.-V., A, C., C, R., *et al.*: Early dropout prediction using data mining: a case study with high school students. Expert Systems **33**(1), 107–124 (2016) https://doi.org/10.1111/exsy.12135

[7] M, S., M, J., O, D., *et al.*: Redefining profit metrics for boosting student retention in higher education. Decision Support Systems **143**, 113493 (2021) https://doi.org/10.1016/J.DSS.2021.113493

[8] S, Z., S, S., P, D., *et al.*: All-year dropout prediction modeling and analysis for university students. Applied Sciences **13**(2), 1143–1143 (2023) https://doi.org/

10.3390/APP13021143

[9] X, X., B, X., X, W., et al.: Review of academic early warning methods based on data mining. Journal of Computer Technology and Electronic Research **1**(3) (2025) https://doi.org/10.70767/JCTER.V1I3.413

[10] X, T., Y, J., H, S.: Hotspots and visual analysis of academic early warning research for college students. Eurasia Journal of Science and Technology **7**(3) (2025) https://doi.org/10.61784/EJST3080

[11] K, L., V, G., W, U.B.: Predicting student dropout: A machine learning approach. European Journal of Higher Education **10**(1), 28–47 (2020) https://doi.org/10.1080/21568235.2020.1718520

[12] K, J., D, M.: Towards predicting student's dropout in university courses using different machine learning techniques. Applied Sciences **11**(7), 3130–3130 (2021) https://doi.org/10.3390/APP11073130

[13] N, M.C.D., O, O.D.L.: A machine learning-based computational system proposal aiming at higher education dropout prediction. Higher Education Studies **10**(4), 12–12 (2020) https://doi.org/10.5539/HES.V10N4P12

[14] F, H., R, L., S, S., et al.: Academic behavior analysis and early warning system based on k-means algorithm. Advances in Computer, Signals and Systems **8**(2) (2024) https://doi.org/10.23977/ACSS.2024.080204

[15] C, S.: Improved fuzzy algorithm for college students' academic early warning. Mathematical Problems in Engineering **2022** (2022) https://doi.org/10.1155/2022/5764800

[16] S, X., Y, Q., H, Y., *et al.*: Integrated computational learning algorithm for undergraduates' academic early warning. Journal of Physics: Conference Series **1815**(1), 012014 (2021) https://doi.org/10.1088/1742-6596/1815/1/012014

[17] H, S., S, F., M, M., *et al.*: Ltboost: A new high-precision method for academic early warning and prediction. Mathematics **14**(9), 1565–1565 (2026) https://doi.org/10.3390/MATH14091565

[18] J, B.: Research on the application of svm-rfe algorithm in the construction of students' academic early warning model. International Journal of Information System Modeling and Design (IJISMD) **16**(1), 1–21 (2025) https://doi.org/10.4018/IJISMD.388927

[19] S, M.W., M, G., S, B.A.: Optimised smote-based imbalanced learning for student dropout prediction. Arabian Journal for Science and Engineering **50**(10), 1–15 (2024) https://doi.org/10.1007/S13369-024-09287-W

[20] N, I.-O., T, B.: Improved student dropout prediction in thai university using ensemble of mixed-type data clusterings. International Journal of Machine Learning and Cybernetics **8**(2), 497–510 (2017) https://doi.org/10.1007/s13042-015-0341-x

[21] C, R., T, B.: Rank biserial stochastic feature embed bivariate kernelized regressive bootstrap aggregative classifier for school student dropout prediction. Concurrency and Computation: Practice and Experience **34**(21) (2022) https://doi.org/10.1002/CPE.7133

[22] P, T., K, S., K, G., *et al.*: Early dropout prediction in moocs through supervised learning and hyperparameter optimization. Electronics **10**(14), 1701–1701 (2021) https://doi.org/10.3390/ELECTRONICS10141701

[23] F, P., H, Z., X, L., *et al.*: Achieving optimal trade-off for student dropout prediction with multi-objective reinforcement learning. PeerJ. Computer science **10**, 2034–2034 (2024) https://doi.org/10.7717/PEERJ-CS.2034

[24] T, Z., H, L., J, T., *et al.*: Enhancing dropout prediction in distributed educational data using learning pattern awareness: A federated learning approach. Mathematics **11**(24), 4977 (2023) https://doi.org/10.3390/MATH11244977

[25] C, C.H., Y, Y.W., H, K.G.: A study on dropout prediction for university students using machine learning. Applied Sciences **13**(21) (2023) https://doi.org/10.3390/APP132112004

[26] Y, W., M, Y., H, H., *et al.*: A study of online academic risk prediction based on neural network multivariate time series features. Concurrency and Computation: Practice and Experience **36**(23), 8251–8251 (2024) https://doi.org/10.1002/CPE.8251

[27] T, K., T, Z., D, N.: Ensemble models based on cnn and lstm for dropout prediction in mooc. Expert Systems With Applications **235** (2024) https://doi.org/10.1016/J.ESWA.2023.121187

[28] N, K., L, G., P, X., *et al.*: Cnn autoencoders and lstm-based reduced order model for student dropout prediction. Neural Computing and Applications **35**(30), 22341–22357 (2023) https://doi.org/10.1007/S00521-023-08894-2

[29] N, K., Z, Y., L, G., et al.: Pmct: Parallel multiscale convolutional temporal model for mooc dropout prediction. Computers and Electrical Engineering **112** (2023) https://doi.org/10.1016/J.COMPELECENG.2023.108989

[30] K, G., S, A., S, A.: Ensemble deep learning network model for dropout prediction in moocs. International journal of electrical and computer engineering systems **14**(2), 187–196 (2023) https://doi.org/10.32985/IJECES.14.2.8

[31] Z, Y., S, Z., D, M., et al.: Mooc dropout prediction using a fusion deep model based on behaviour features. Computers and Electrical Engineering **104** (2022) https://doi.org/10.1016/J.COMPELECENG.2022.108409

[32] P, F., H, B., Z, C., *et al.*: A survival analysis based volatility and sparsity modeling network for student dropout prediction. PloS one **17**(5), 0267138–0267138 (2022) https://doi.org/10.1371/JOURNAL.PONE.0267138

[33] H, W.S., M, K.J., D, K., et al.: University student dropout prediction using pretrained language models. Applied Sciences **13**(12) (2023) https://doi.org/10.3390/APP13127073

[34] Y, Q., Z, S., L, L.: Research on the academic early warning model of distance education based on student behavior data in the context of covid-19. International Journal of Advanced Computer Science and Applications (IJACSA) **13**(10) (2022) https://doi.org/10.14569/IJACSA.2022.0131074

[35] X, X., Q, W.: Driving stem learning effectiveness: dropout prediction and intervention in moocs based on one novel behavioral data analysis approach. Humanities and Social Sciences Communications **11**(1) (2024) https://doi.org/10.1057/S41599-024-02882-0

[36] M, Z., M, M., B, C., *et al.*: A machine learning based model for student's dropout prediction in online training. Education and Information Technologies **29**(12), 15793–15812 (2024) https://doi.org/10.1007/S10639-024-12500-W

[37] X, X., Q, W.: Dropout prediction and decision feedback supported by multi temporal sequences of learning behavior in moocs. International Journal of Educational Technology in Higher Education **20**(1) (2023) https://doi.org/10.1186/S41239-023-00400-X

[38] Z, Y., Z, J., Z, J.: Prediction of learners' dropout in e-learning based on the unusual behaviors. Interactive Learning Environments **31**(3), 1796–1820 (2023) https://doi.org/10.1080/10494820.2020.1857788

[39] L, H., C, X., Z, F.: Learning behavior feature fused deep learning network model for mooc dropout prediction. Education and Information Technologies **29**(3), 3257–3278 (2023) https://doi.org/10.1007/S10639-023-11960-W

[40] C, J.: Dropout prediction model in mooc based on clickstream data and student sample weight. Soft Computing **25**(14), 1–18 (2021) https://doi.org/10.1007/S00500-021-05795-1

[41] B, B.R., J, C., D, T.: Dropout prediction in moocs using deep learning and machine learning. Education and Information Technologies **27**(8), 11499–11513 (2022) https://doi.org/10.1007/S10639-022-11068-7

[42] S, N., S, U.: Mooc dropout prediction using fiar-ann model based on learner behavioral features. International Journal of Advanced Computer Science and Applications (IJACSA) **13**(9) (2022) https://doi.org/10.14569/IJACSA.2022.0130972

[43] M, A.A., C, H., H, M.I.: Deep analytic model for student dropout prediction in massive open online courses. Computers and Electrical Engineering **93** (2021) https://doi.org/10.1016/J.COMPELECENG.2021.107271

[44] S, G., A, S.S.: Deciphering the attributes of student retention in massive open online courses using data mining techniques. Education and Information Technologies **24**(3), 1973–1994 (2019) https://doi.org/10.1007/s10639-018-9829-9

[45] A, B., F, B.Z., H, H.: A modular and explainable machine learning pipeline for student dropout prediction in higher education. Algorithms **18**(10), 662–662 (2025) https://doi.org/10.3390/A18100662

[46] K, C.G.J., B, S.D.A., B, P.J.: An explainable machine learning approach for student dropout prediction. Expert Systems With Applications **233** (2023) https://doi.org/10.1016/J.ESWA.2023.120933

[47] N, M., M, R.: Interpretable dropout prediction: Towards xai-based personalized intervention. International Journal of Artificial Intelligence in Education **34**(2), 274–300 (2023) https://doi.org/10.1007/S40593-023-00331-8

[48] X, Y.: Optimization and implementation of management technology integrated with data analysis for college students' course evaluation and academic early warning. Systems and Soft Computing **7**, 200255–200255 (2025) https://doi.org/10.1016/J.SASC.2025.200255

[49] A, R.M., L, Z.E.: A model for predicting dropout of higher education students. Data Science and Management **8**(1), 72–85 (2025) https://doi.org/10.1016/J.DSM.2024.07.001

[50] M, B., S, A.: A novel approach based on fuzzy rule and lsowl–cnn forecasting students with dropout prediction and recommendation model. Wireless Personal Communications **136**(1), 61–80 (2024) https://doi.org/10.1007/S11277-024-11068-5

[51] Q, M.E., B, F.M.M., P, R.V., *et al.*: Early prediction of at-risk students in secondary education: A countrywide k-12 learning analytics initiative in uruguay. Information **13**(9), 401–401 (2022) https://doi.org/10.3390/INFO13090401