INTERPRETABLE MULTI-MODEL FUSION FOR ACADEMIC PERFORMANCE PREDICTION

Shen Mingzhu

"Tashkent Institute of Irrigation and Agricultural Mechanization Engineers" National Research University 100000,39, Kori Niyazi street,Tashkent,Republic of Uzbekistan
https://orcid.org/0009-0002-0577-8156
E-mail:18265987700@163.com

KEYWORDS

academic performance prediction, multi-model fusion, ensemble learning, learning analytics, explainable artificial intelligence, higher education

ABSTRACT

Academic performance prediction is a core task in educational data mining and learning analytics, and its outcomes directly inform academic intervention decisions in higher education. A single model can hardly exploit multi-source, heterogeneous learning-behavior features; this study therefore proposes a multi-model fusion framework in which logistic regression, random forest, XGBoost, LightGBM and CatBoost serve as heterogeneous base learners, and decision-level fusion is realized through a stacking (logistic regression) meta-learner and an MLP (neural network) meta-learner. Stratified five-fold cross-validation on the public OULAD dataset (32,593 course-enrollment records, 46 features) shows that the best single model, CatBoost, attains an AUC of 0.9730, while both fusion strategies reach 0.9727, comparable to the best single model and achieving a better recall-precision balance. Ablation results show that assessment scores are the dominant information source, while virtual learning environment (VLE) behavioral features provide a stable incremental gain; SHAP analysis identifies the number of assessments attempted, mean score and VLE active days as the most important predictors. When strong single models are already near the performance ceiling, multi-model fusion thus adds value chiefly by reducing model-selection risk and by providing interpretable evidence for prediction.

1 Introduction

Student attrition and academic failure in higher education directly affect institutional sustainability and educational equity. Learning analytics (LA) and educational data mining (EDM) now make it possible to predict academic outcomes from digital footprints [1,2]. Behavioral logs recorded in virtual learning environments (VLEs), such as visits, submissions and time on task, together with early grades and demographic features, can support proactive academic prediction and intervention systems.

Among modeling approaches, ensemble learning has been widely shown to outperform single algorithms. Gradient boosting models such as XGBoost and LightGBM often outperform bagging models such as random forest in various educational scenarios [3,4]; stacking feeds the outputs of heterogeneous base models into a meta-learner through a two-layer structure, further improving performance [5,6].

However, the literature still has several gaps. First, most studies rely on a single data source or a single family of models, lacking systematic fusion of heterogeneous models. Second, existing fusion weights are mostly fixed and do not adapt to sample characteristics. Third, fused models are often insufficiently interpretable to be translated into actionable evidence for educators. Systematic reviews likewise note that classification methods dominate, while multi-source data integration and hybrid or ensemble modeling remain underexplored [1,7].

To address these gaps, we construct and empirically validate a multi-model fusion framework for academic performance prediction, in which heterogeneous base models are fused through a learnable MLP (neural network) meta-learner that nonlinearly combines base-model outputs, together with explainability analysis. The contributions are: (1) a fusion architecture of heterogeneous base learners covering linear, bagging and multiple boosting families, with

(cc) BY

reserved interfaces for temporal models (LSTM/Transformer) and graph models (GCN/GAT); (2) a nonlinear fusion mechanism based on an MLP meta-learner that dynamically combines base-model outputs on a per-sample basis; (3) empirical validation on a public higher-education dataset with SHAP-based explainability analysis.

2 Theoretical Background

2.1 Educational Data Mining and Learning Analytics

EDM and LA use multi-source data from institutional information systems, including grades, behavioral logs and demographics, to model academic outcomes and support early intervention and evidence-based decision-making [7,8].

2.2 Principles of Ensemble Learning

Ensemble learning improves generalization by combining multiple base learners to reduce variance or bias, and mainly comprises three paradigms: bagging, represented by random forest [9]; boosting, represented by XGBoost, LightGBM and CatBoost [4,10]; and stacking, which takes the outputs of base models as the inputs of a meta-learner [11,12].

2.3 Multi-Model Fusion and Meta-Learning

Multi-model fusion can be divided into early (feature-level) fusion and late (decision-level) fusion. Late fusion combines the predictions of base models through voting, weighted averaging or a meta-learner (LR/MLP/GBM). Building on this, we introduce a learnable nonlinear combination implemented by an MLP meta-learner, so that the fusion weights are no longer fixed. Recent studies also show that multi-model fusion can achieve performance comparable to or better than single models in academic performance prediction in higher education [13].

2.4 Temporal Modeling and Graph Neural Networks (Framework Extensions)

As extensible interfaces of the framework, LSTM and Transformer models can characterize the temporal evolution of grades and VLE behavior, while graph neural networks (GCN/GAT) can model peer and course-enrollment relationships among students. The empirical part of this study focuses on heterogeneous tabular base learners; temporal and graph models are discussed as future extensions.

2.5 Explainable Artificial Intelligence

SHAP [14] gives instance-level and global explanations of fused models, helping educators understand key factors [3,15] and supporting accountability in academic prediction and intervention systems. Incorporating educational domain knowledge into explainable models has also been shown to further improve the usability and actionability of predictions [16].

3 Materials and Methods

This section describes the data source and prediction target (3.1), feature engineering (3.2), base models (3.3), fusion strategies (3.4), evaluation metrics and validation protocol (3.5), experimental design (3.6) and experimental environment (3.7). The experiments center on five heterogeneous base learners and two decision-level fusion strategies, adopt stratified five-fold cross-validation with out-of-fold (OOF) probabilities to prevent data leakage, quantify the contribution of each feature source through ablation, and identify key predictors with SHAP.

3.1 Data Source

This study uses the public Open University Learning Analytics Dataset (OULAD) [17,18] from higher education for empirical validation. The dataset is released by the UK Open University with an assigned DOI (10.1038/sdata.2017.103), is publicly available and version-fixed, which simplifies reproduction. Prior work has systematically compared supervised models for early risk identification on OULAD [18], providing a baseline reference for our fusion design. The dataset contains 32,593 course-enrollment records covering demographics, assessment scores and detailed VLE interaction logs, and is a mainstream benchmark for academic performance prediction

(cc) BY

in higher education. Following common practice in OULAD studies [17,18], the prediction target is a binary final outcome: the positive class comprises Pass or Distinction, and the negative class comprises Fail or Withdrawn (withdrawal is regarded as a form of not attaining the academic outcome), with a positive rate of 47.2%. Should institutional data from the authors' university (TIIAME) be used in the future, ethics and data-use approval must be obtained beforehand.

3.2 Data Preprocessing and Feature Engineering

Taking one course enrollment (code_module, code_presentation, id_student) as the data unit, we construct 46 features: static features (demographics, number of previous attempts, registered credits, etc.), assessment features (number of assessments attempted, mean score, first score, mean of the first two assessments, mean submission delay, etc.) and VLE behavioral features (total clicks and active days, aggregated per student from more than 10.65 million VLE records). Categorical features are one-hot encoded, and missing numerical values are filled with 0. The assessment and VLE behavioral features are aggregated over the entire course period (including all assessment records and all VLE interaction logs); the prediction setting is therefore retrospective, and the mean of the first two assessments serves as a proxy feature for early prediction. For early-warning deployment during the course, features must be re-truncated according to the intervention time window (see the limitations in Section 5).

3.3 Base Model Design

Five heterogeneous base learners are used: logistic regression (linear baseline), random forest (bagging), and XGBoost, LightGBM and CatBoost (three boosting families), covering the linear, bagging and boosting paradigms to ensure diversity. Logistic regression, random forest, LightGBM and CatBoost all adopt class weights (balanced) to correct the slight difference in class priors; the key hyperparameters are listed in Table 1. The framework reserves extension interfaces at the base-learner level (see Section 2.4).

Table 1

Key hyperparameters of the base models (unlisted items use library defaults; random seeds are all 42)

Model	Key hyperparameters
Logistic Regression	C=0.5, max_iter=2000, class_weight=balanced
Random Forest	n_estimators=300, max_depth=None, class_weight=balanced
XGBoost	n_estimators=400, max_depth=5, learning_rate=0.05, subsample=0.9, colsample_bytree=0.9, eval_metric=logloss
LightGBM	n_estimators=400, max_depth=6, learning_rate=0.05, subsample=0.9, colsample_bytree=0.9, class_weight=balanced
CatBoost	iterations=400, depth=6, learning_rate=0.05, auto_class_weights=Balanced

3.4 Fusion Strategies

Two decision-level fusion strategies are adopted: (a) stacking, which trains a logistic-regression meta-learner on the out-of-fold (OOF) predicted probabilities of the base models; and (b) MLP meta-learner fusion, in which an MLP meta-learner performs a learnable nonlinear combination of the base models' OOF probabilities, so that the fusion output adapts to each sample's base-model prediction pattern. The MLP meta-learner has two hidden layers (32 and 16 units, ReLU activation), uses the Adam optimizer with L2 regularization (alpha=1e-3), a maximum of 2,000 iterations and a fixed random seed of 42; its input is the OOF probabilities of the five base models, and its predictions are likewise generated through stratified five-fold cross-validation to avoid meta-level overfitting. Soft voting and performance-based static weighting serve as comparison baselines. The overall fusion framework is shown in Figure 1.

(cc) BY

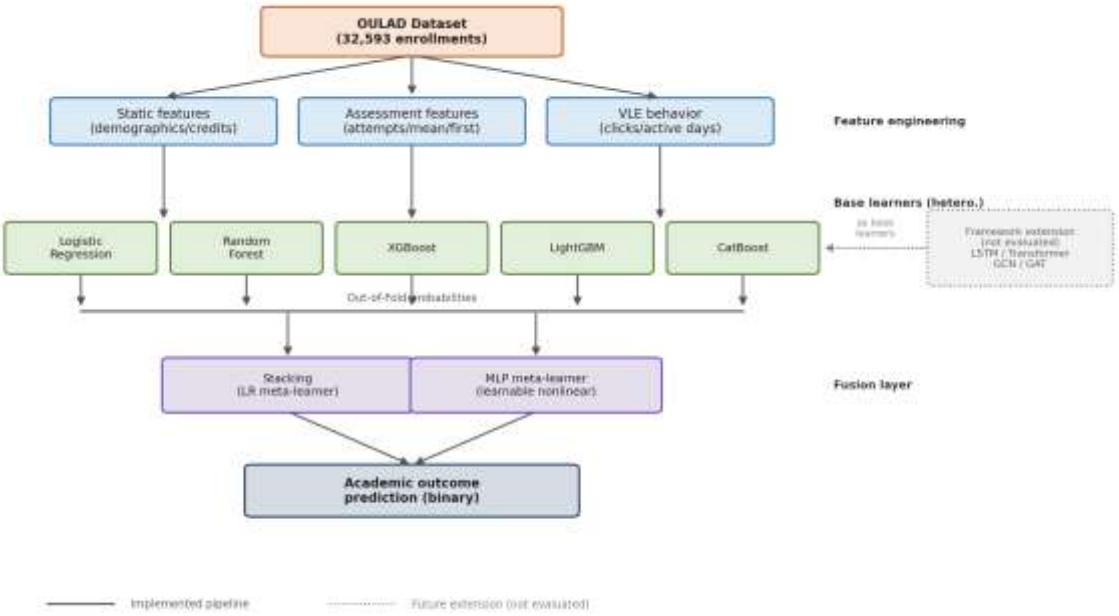


Figure 1 The multi-model fusion framework

Figure 1 The multi-model fusion framework: OULAD data -> three-source feature engineering -> heterogeneous base learners -> OOF predicted probabilities -> stacking/MLP meta-learner fusion -> academic outcome prediction (binary). Solid lines denote the pipeline implemented in this study; the dashed box on the right denotes temporal and graph models (LSTM/Transformer, GCN/GAT) that can be incorporated as base learners in the future and are not empirically evaluated here.

3.5 Evaluation Metrics and Validation Protocol

Stratified five-fold cross-validation is used, and OOF probabilities are supplied to the fusion layer to prevent data leakage. The metrics are Accuracy, Precision, Recall, F1 and AUC-ROC; SHAP is used for global explainability analysis.

3.6 Experimental Design

(1) Comparison of base models and fusion strategies; (2) ablation with three feature settings: static only, static plus assessment, and full features (including VLE); (3) identification of key factors via SHAP.

3.7 Experimental Environment and Reproducibility

All randomized steps use a fixed random seed of 42 (including the StratifiedKFold data splits and model initialization) to ensure reproducibility. The official OULAD release snapshot is the sole data source. The runtime environment is Python 3.13.12, with core dependency versions: numpy 2.5.1, pandas 3.0.5, scikit-learn 1.9.0, XGBoost 3.4.1, LightGBM 4.7.0, CatBoost 1.2.10, SHAP 0.52.0 and SciPy 1.18.0. In multithreaded CatBoost and LightGBM, floating-point summation order may shift individual metrics by plus or minus 1 in the fourth decimal place. The feature-engineering and fusion code is publicly available at https://github.com/smzmria-gm/oulad-multimodel-fusion, together with the dependency list (requirements.txt), experiment logs and result files, which reproduces all numerical results reported here.

4 Results

4.1 Data Description and Baseline Performance

The dataset contains 32,593 enrollment records, with 15,385 positive samples (Pass/Distinction, 47.2%) and 17,208 negative samples (52.8%), an approximately balanced class distribution. Feature engineering yields a 46-dimensional feature matrix. Table 2 reports the mean performance of each base model under five-fold cross-validation.

(cc) BY

Table 2

Mean performance of the base models under stratified five-fold cross-validation (values rounded to four decimal places).

Model	Accuracy	F1	AUC	Recall	Precision
Logistic Regression	0.8831	0.8792	0.9501	0.9012	0.8583
Random Forest	0.9080	0.9068	0.9688	0.9479	0.8692
XGBoost	0.9142	0.9119	0.9722	0.9410	0.8845
LightGBM	0.9136	0.9116	0.9719	0.9446	0.8809
CatBoost	0.9155	0.9139	0.9730	0.9497	0.8807

The standard deviations of all base models across the five folds do not exceed 0.0048 (AUC standard deviation <= 0.0022), indicating stable and reproducible results.

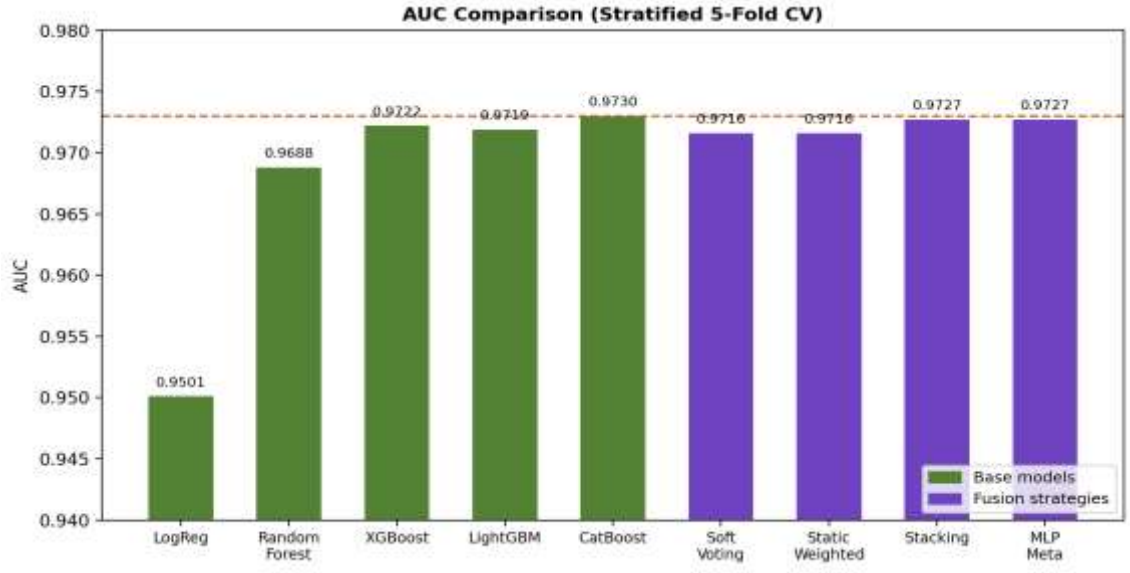
4.2 Performance of the Fusion Framework

Table 3 compares the four fusion strategies with the best single model (see also Figure 2). Static weighting assigns weights via softmax normalization of the base models' OOF AUCs; since the base models perform similarly, the normalized weights are nearly uniform, and its results are essentially identical to soft voting. The AUC difference between the stacking/MLP meta-learner fusion (0.9727) and the best single model CatBoost (0.9730) is only 0.0003, within the range of cross-validation fluctuation. The three are tied on Accuracy (0.9155 / 0.9150 / 0.9155), while F1 is slightly lower than that of CatBoost (0.9129 / 0.9128 vs. 0.9139); however, recall and precision are better balanced (stacking 0.9383/0.8889, MLP meta-learner fusion 0.9423/0.8851), and both clearly outperform the logistic-regression baseline.

Table 3

Comparison of fusion strategies with the best single model (OOF evaluation, rounded to four decimal places). The dashed line in Figure 2 marks the AUC of the best single model, CatBoost (0.9730).

Strategy	Accuracy	F1	AUC	Recall	Precision
Soft Voting	0.9144	0.9128	0.9716	0.9492	0.8790
Static Weighting	0.9144	0.9128	0.9716	0.9493	0.8790
Stacking(LR)	0.9155	0.9129	0.9727	0.9383	0.8889
MLP Meta-Learner Fusion	0.9150	0.9128	0.9727	0.9423	0.8851
Best Single (CatBoost)	0.9155	0.9139	0.9730	0.9497	0.8807



(cc) BY

Figure 2 AUC comparison of the models (stratified five-fold cross-validation). Green bars denote base models and purple bars denote fusion strategies; the dashed line marks the AUC of the best single model, CatBoost (0.9730).

4.3 Ablation and the Contribution of MLP Meta-Learner Weighting

The ablation results (Table 4 and Figure 3) show the marginal contributions of feature sources: with static demographic features only, the fusion AUC is merely 0.6565; adding assessment features raises it to 0.9696, and adding VLE behavioral features brings it to 0.9727. Assessment scores are thus the dominant information source, while VLE logs add a smaller but steady gain. As for fusion schemes, stacking and MLP meta-learner fusion outperform fixed-weight fusion by only 0.0011 in AUC, so the accuracy gain of adaptive weighting on this dataset is limited.

Table 4

Ablation: fusion AUC under different feature settings (stacking evaluation).

Feature Setting	Fusion AUC
Static only	0.6565
Static + Assessment	0.9696
Full (incl. VLE)	0.9727

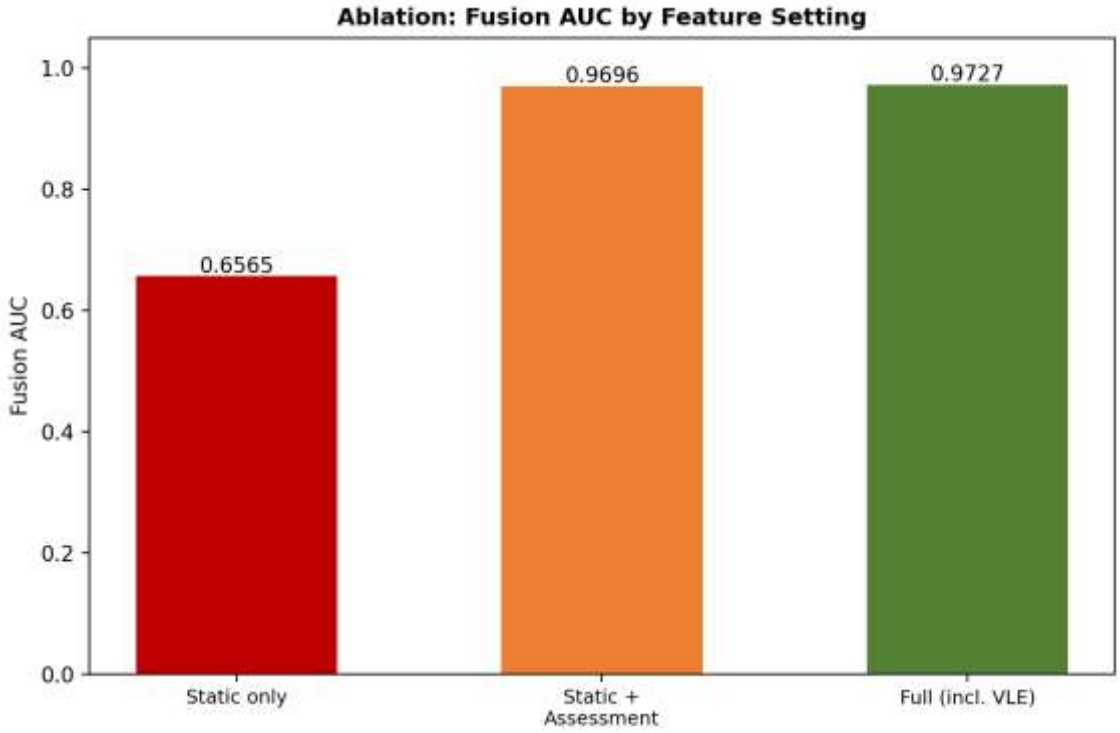


Figure 3 Ablation: fusion AUC under different feature settings. The AUC is only 0.6565 with static features, rises to 0.9696 after adding assessment scores, and reaches 0.9727 after further adding VLE behavioral features.

4.4 Interpretability and Key Factors

A SHAP global analysis of the XGBoost base model ranks the top ten features by mean absolute SHAP value (Figure 4): number of assessments attempted, mean score, VLE active days, first score, mean submission delay, VLE total clicks, highest education (lower than A Level), registered credits, mean of the first two assessments, and registration date. The number of assessments attempted and the mean score contribute most prominently, and VLE behavioral features (active days, total clicks) also rank high, in line with the established educational intuition that engagement shapes academic outcomes.

(cc) BY

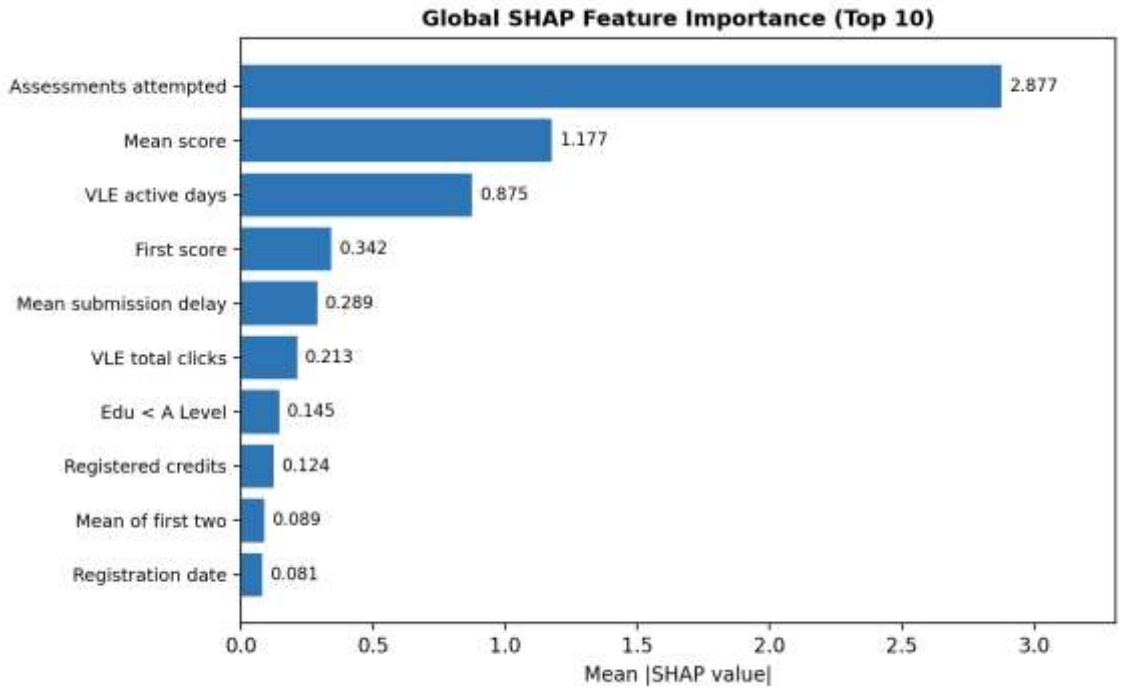


Figure 4 Global SHAP feature importance of the XGBoost base model (top ten). The horizontal axis shows the mean absolute SHAP value, reflecting each feature's average marginal contribution to model predictions.

5 Discussion

At the single-model level, gradient boosting models (AUC of roughly 0.972-0.973) outperform random forest (0.969) and logistic regression (0.950), agreeing with the superiority of boosting models reported in previous studies [4]. At the fusion level, the stacking/MLP meta-learner fusion AUC of our framework (0.9727) is comparable to the accuracy levels reported in recent ensemble studies [6,19].

As shown in Section 4.2, fusion offers no significant accuracy advantage over the best single model, CatBoost (0.9730). This echoes the conclusion of Kelicha et al. [20] that ensemble gains are "modest and context-dependent": on this dataset, strong boosting models are already near the performance ceiling. The significance of fusion therefore lies in three aspects: (1) reducing model-selection risk, since fusion does not rely on the inductive bias of a single model; (2) improving decision balance, as MLP meta-learner fusion achieves a better recall-precision trade-off (Recall 0.9423 / Precision 0.8851), fitting the intervention goal of "missing as few at-risk students as possible"; and (3) interpretability, as SHAP yields key factors that are consistent across models and easy to adopt in teaching practice.

The ablation results reveal a hierarchy of feature value: assessment scores are the dominant signal, and VLE behavior provides a stable marginal gain. This suggests that academic prediction and intervention systems in higher education should prioritize connecting assessment data streams and incorporate VLE logs where available. Future work includes incorporating temporal models (LSTM/Transformer) and graph models (GCN/GAT) into empirical evaluation to capture dynamic and relational information, conducting external validation on institutional data from the authors' university (TIIAME), and analyzing prediction time windows for different course modules.

This study has the following limitations: (1) only the single OULAD dataset is used, and the generalization of the conclusions to other institutions and curricula is untested; (2) only decision-level fusion is implemented, and temporal and graph models have not been empirically evaluated; (3) the differences among fusion strategies are small, and their statistical significance remains to be established; (4) the features are aggregated over the entire course period and include final assessment scores, which is a retrospective setting, so the conclusions do not directly generalize to in-course early warning;

(cc) BY

deployment in that scenario requires re-truncating the features along the chosen time window.

6 Conclusion

This study proposes and empirically validates an interpretable multi-model fusion framework for academic performance prediction, which realizes decision-level fusion by combining five heterogeneous base learners with stacking and MLP meta-learners. Stratified five-fold cross-validation on OULAD (32,593 enrollment records) shows that the fusion framework attains an AUC of 0.9727, on par with the best single model, CatBoost, and better balanced between recall and precision. Ablation and SHAP analyses identify assessment scores and learning engagement as the dominant predictors, with VLE behavior adding a smaller but steady contribution. Where strong boosting models are already near their accuracy limit, fusion adds little raw accuracy; its value lies in robustness to model choice, a better recall-precision balance, and cross-model agreement on key factors. Future work includes external validation on institutional data, incorporating temporal and graph models into empirical evaluation, and reconstructing features according to time windows for early-warning scenarios.

References

1. Abu Saa A., Al-Emran M., Shaalan K. Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques // Technology, Knowledge and Learning. 2019. Vol. 24. Pp. 567–598. DOI: 10.1007/s10758-019-09408-7.
2. Alhazmi E., Sheneamer A. Early predicting of students' performance in higher education // IEEE Access. 2023. Vol. 11. Pp. 27579–27589. DOI: 10.1109/ACCESS.2023.3250702.
3. Arévalo-Cordovilla F. E., Peña M. Evaluating ensemble models for fair and interpretable prediction in higher education using multimodal data // Scientific Reports. 2025. DOI: 10.1038/s41598-025-15388-9.
4. Mahawar K., Rattan P. Empowering Education: Harnessing Ensemble Machine Learning Approach and ACO-DT Classifier for Early Student Academic Performance Prediction // Education and Information Technologies. 2025. Vol. 30. No. 4. Pp. 4639–4667. DOI: 10.1007/s10639-024-12976-6.
5. Li L. A Hybrid Ensemble Learning Framework for Educational Data Mining and Learning Analytics Research from an Interdisciplinary Perspective // Journal of Computing & Biomedical Informatics. 2026. Vol. 11. No. 2. DOI: 10.56979/1102/2026/1467.
6. Hermaliani E. H., Fanani A. Z., Santoso H. A., Affandy. An Explainable Stacking Ensemble Framework for Student Academic Performance Prediction in Learning Management Systems // Ingénierie des systèmes d information. 2026. Vol. 31. No. 4. Pp. 1251–1265. DOI: 10.18280/isi.310420.
7. Aldisa R. T., Rochim A. F., Triayudi A. Student Achievement Prediction Models: A PRISMA-Based Systematic Literature Review // Journal of Information Systems and Informatics. 2026. Vol. 8. No. 2. Pp. 2638–2663. DOI: 10.63158/journalisi.v8i2.1526.
8. Okoye K., Nganji J. T., Escamilla J., Hosseini S. Machine Learning Model (RG-DMML) and Ensemble Algorithm for Prediction of Students' Retention and Graduation in Education // Computers and Education: Artificial Intelligence. 2024. Vol. 6. Pp. 100205. DOI: 10.1016/j.caeai.2024.100205.
9. Breiman L. Random forests // Machine Learning. 2001. Vol. 45. No. 1. Pp. 5–32. DOI: 10.1023/A:1010933404324.
10. Chen T., Guestrin C. XGBoost: a scalable tree boosting system // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. Pp. 785–794. DOI: 10.1145/2939672.2939785.
11. Wolpert D. H. Stacked generalization // Neural Networks. 1992. Vol. 5. No. 2. Pp. 241–259. DOI: 10.1016/S0893-6080(05)80023-1.
12. Predicting student academic achievement using stacked ensemble learning with deep neural networks and fuzzy-based feature selection // Scientific Reports. 2025. DOI: 10.1038/s41598-025-20779-z.

(cc) BY

13. Zou W., Zhong W., Du J., Yuan L. Prediction of Student Academic Performance Utilizing a Multi-Model Fusion Approach in the Realm of Machine Learning // Applied Sciences. 2025. Vol. 15. No. 7. Pp. 3550. DOI: 10.3390/app15073550.
14. Lundberg S. M., Lee S.-I. A unified approach to interpreting model predictions // Advances in Neural Information Processing Systems 30. 2017. Pp. 4765–4774.
15. Lu Y., Rajendra A. S., Zhang J., Zhao T. Explainable Machine Learning for Student Performance Prediction // AI in Education. 2026. Vol. 2. No. 2. Pp. 17. DOI: 10.3390/aieduc2020017.
16. Qiang M., Liu Z., Zhang R. Explainable AI in education: integrating educational domain knowledge into the deep learning model for improved student performance prediction // Scientific Reports. 2026. Vol. 16. No. 1. DOI: 10.1038/s41598-026-40538-y.
17. Kuzilek J., Hlosta M., Zdrahal Z. Open University Learning Analytics dataset // Scientific Data. 2017. Vol. 4. Pp. 170171. DOI: 10.1038/sdata.2017.103.
18. Omarbekova A., Nazyrova A., Bekmanova G., Lamasheva Z., Turebayeva R., Abzal Z. Comparative Evaluation of Machine Learning Models for Early Prediction of Student Academic Performance Using the Open University Learning Analytics Dataset // International Journal of Interactive Mobile Technologies (iJIM). 2026. Vol. 20. No. 17. DOI: 10.3991/ijim.v20i17.62574.
19. Kalani N., Joshi H. An Improved Hybrid Stacking Ensemble Model for Predicting Student Academic Performance Using Machine Learning // International Journal for Research in Applied Science and Engineering Technology. 2026. Vol. 14. No. 3. Pp. 318–321. DOI: 10.22214/ijraset.2026.77793.
20. Kelicha O., Blessing U. F., Anyanwu A., Nwagbara C. Ensemble Machine Learning for Student Performance Prediction in Higher Education: A Critical Review of Predictive Gains, Methodological Quality and Deployment Constraints // Asian Journal of Research in Computer Science. 2026. Vol. 19. DOI: 10.9734/ajrcos/2026/v19i9911.

(cc) BY *dtai.uz*