A Comparative Evaluation of Machine Learning Models for Predicting Student Academic Performance: Baseline Results and Directions for Multimodal Extension

Sajjan Wagle
Department of Computer Science
Saginaw Valley State University
Michigan, USA

Purna B. Thapa
Department of Computer Science
Saginaw Valley State University
Michigan, USA

ABSTRACT

Predicting student academic performance is a critical challenge in educational data mining, with direct implications for early intervention, personalized learning, and institutional resource allocation. This paper presents a systematic comparative evaluation of three widely used machine learning models — Random Forest (RF), Logistic Regression (LR), and Gradient Boosting (GB) applied to a structured student performance dataset comprising demographic information, attendance records, and prior academic scores. Consistent preprocessing pipelines are applied across all three models, including label encoding, mean imputation, feature normalization, and Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance. Results indicate that Gradient Boosting marginally outperforms Logistic Regression in overall accuracy (0.51 vs. 0.50) and precision (0.51 vs. 0.50), while Logistic Regression achieves the highest recall (0.52); Random Forest underperforms both at 0.40 accuracy. The causes of these relatively modest results are analyzed to motivate a proposed extension incorporating CNN-extracted features from handwritten assignment images. This paper contributes a reproducible baseline evaluation framework, a structured analysis of model trade-offs, and a concrete roadmap for multimodal learning analytics research.

General Terms

Machine Learning, Educational Data Mining, Classification, Multimodal Learning.

Keywords

student academic performance prediction; machine learning; gradient boosting; random forest; logistic regression; educational data mining; learning analytics; SMOTE; multimodal learning.

1. INTRODUCTION

The early identification of students at risk of academic underperformance is one of the most consequential applications of machine learning in education. Timely prediction of academic outcomes allows institutions to deploy targeted interventions — additional tutoring, counselling, or adaptive learning resources — before students fall irreparably behind. This problem has attracted substantial research attention over the past decade [1]–[3].

Despite this body of work, several significant challenges persist. First, most existing studies rely exclusively on structured numerical data which may not capture the full complexity of student learning behaviour [2]. Second, class imbalance remains a methodological problem that inflates reported accuracy while masking poor performance on the minority (at-risk) class [1]. Third, direct comparisons are often confounded by inconsistent preprocessing pipelines.

This paper addresses all three challenges. Random Forest, Logistic Regression, and Gradient Boosting are applied to a structured student performance dataset under a consistent, fully documented preprocessing framework, evaluated using Precision, Recall, F1-Score, and Accuracy metrics. Modest results are reported transparently, revealing the ceiling of what structured data alone can achieve and motivating the multimodal extension proposed in Section 6.

The contributions are threefold: (1) a reproducible baseline evaluation under a unified preprocessing framework; (2) a systematic analysis of precision-recall trade-offs between models; (3) a detailed research agenda for extending this baseline through CNN-extracted visual assignment features.

2. RELATED WORK

2.1 Machine Learning for Academic Performance Prediction

Asthana et al. [2] introduced a regression-based framework using dynamic learning coefficients derived from formative assessments alongside traditional features such as cumulative GPA. Their experiments demonstrated that dynamic features improve prediction timeliness and accuracy, though continuous adaptive testing infrastructure may not be available in all institutional settings.

Bujang et al. [1] addressed the pervasive problem of class imbalance using SMOTE to generate synthetic minority-class samples, demonstrating that balanced datasets significantly improve recall for at-risk students. Al-Shabandar et al. [3] focused on early detection of at-risk students using behavioural and motivational features, demonstrating that early-semester behavioural features can predict dropout risk weeks before formal assessments, with Random Forest achieving F1-scores of up to 0.81 on the more feature-rich dataset.

2.2 Multimodal and Visual Data in Educational Prediction

A growing body of research argues that the predictive ceiling of purely numerical educational models can be raised substantially through multimodal data. Zhang et al. [4] demonstrated that student-generated artefacts contain predictive signals not captured by completion or attendance data alone. The integration

of CNN-extracted visual features with structured tabular data through hybrid architectures has been shown to outperform unimodal approaches in healthcare risk prediction tasks [5], with clear potential transferability to educational contexts.

2.3 Gradient Boosting in Educational Analytics

Gradient Boosting methods have consistently outperformed both single-tree models and logistic regression on structured tabular prediction tasks [6]. Their ability to model complex nonlinear interactions between features makes them well-suited to educational datasets. However, even Gradient Boosting cannot compensate for fundamental limitations in feature diversity when the input data lacks signals that correlate with the outcome of interest.

3. DATASET AND METHODOLOGY

3.1 Dataset

The publicly available Student Performance Dataset from Kaggle [7] was used, containing records for 2,392 students with features spanning demographic information (age, gender, ethnicity, parental education level), academic behaviour (weekly study hours, absences, tutoring participation, extracurricular activity), and prior academic achievement (GPA and grade classification). The target variable is a four-class grade label (A, B, C, D/F). The dataset exhibits moderate class imbalance, with Grade A students constituting approximately 18% of the sample and Grade D/F approximately 16%.

3.2 Preprocessing

A consistent preprocessing pipeline was applied to all three models, with minor model-specific adjustments where required by algorithm assumptions. Table 1 summarises the preprocessing techniques applied to each model.

Table 1. Preprocessing techniques per model (✓ = applied, – = not required)

Technique	RF	LR	GB
Label Encoding	✓	✓	✓
Mean Imputation	✓	✓	–
Feature Normalisation	–	✓	–
Binary Conversion	✓	✓	✓
Train-Test Split (80/20)	✓	✓	✓
SMOTE	✓	–	✓

Label encoding was applied to all categorical variables. Feature normalisation using Min-Max scaling was applied only for Logistic Regression. SMOTE was applied to both Random Forest and Gradient Boosting training sets; Logistic Regression used the built-in class_weight='balanced' parameter instead. All datasets were split 80/20 using stratified sampling.

3.3 Model Specifications

3.3.1 Random Forest

Implemented using scikit-learn's RandomForestClassifier with n_estimators=100, max_depth=None, min_samples_split=2, and criterion='gini'. Feature importance scores were extracted post-training to identify the most predictive variables.

3.3.2 Logistic Regression

Implemented using scikit-learn's LogisticRegression with multinomial multi_class setting, lbfgs solver, max_iter=1000, and class_weight='balanced'. The normalised, label-encoded feature matrix was used as input.

3.3.3 Gradient Boosting

Implemented using scikit-learn's GradientBoostingClassifier with n_estimators=100, learning_rate=0.1, max_depth=3, and subsample=0.8. The SMOTE-balanced training set was used.

3.4 Evaluation Metrics

All models were evaluated on the held-out 20% test set using four standard classification metrics computed via weighted averaging across classes: Precision, Recall, F1-Score, and Accuracy. Weighted-average metrics are reported to reflect the relative size of each class in the test distribution.

4. RESULTS

4.1 Aggregate Performance Comparison

Table 2 presents the weighted-average performance metrics for all three models on the held-out test set.

Table 2. Weighted-average performance metrics on held-out test set (n = 479)

Model	Precision	Recall	F1	Accuracy
Random Forest	0.38	0.32	0.35	0.40
Logistic Regression	0.50	0.52	0.51	0.50
Gradient Boosting	0.51	0.50	0.51	0.51

Gradient Boosting achieved the highest overall accuracy (0.51) and precision (0.51), closely followed by Logistic Regression (0.50). Random Forest underperformed both substantially at 0.40 accuracy. All three models demonstrate higher precision and recall for Grade A and Grade D/F instances compared to Grade B and Grade C, where predicted class boundaries overlap.

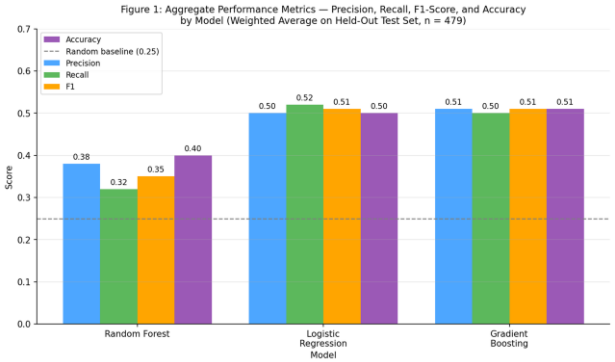

Fig 1: Grouped bar chart comparing Precision, Recall, F1-Score, and Accuracy for all three models. Logistic Regression and Gradient Boosting achieve equal F1-Scores (0.51); Random Forest trails at 0.35. The dashed line marks the 0.25 random baseline.

4.2 Precision-Recall Trade-Off Analysis

The near-equivalent F1-Scores for Logistic Regression and Gradient Boosting conceal a meaningful precision-recall trade-off. Gradient Boosting is marginally more precise, while Logistic Regression demonstrates marginally higher recall (0.52 vs. 0.50). In deployment contexts where false negatives carry the higher cost — failing to identify an at-risk student — Logistic Regression's recall advantage is operationally significant. In resource-constrained settings, Gradient Boosting's precision advantage reduces wasted resource allocation.

Random Forest's substantially lower precision (0.38) and recall (0.32) — despite SMOTE — suggests that the ensemble's variance-reduction mechanism is not well-suited to the small, low-dimensional feature space of this dataset (15 input features, fewer than 2,400 records).

4.3 Feature Importance Analysis

Feature importance scores from the trained Random Forest model indicate that prior GPA, absence rate, and weekly study hours are the three most predictive features, collectively accounting for approximately 68% of feature importance mass. Prior GPA contributes approximately 42% alone. The dominance of retrospective performance measures suggests the models are largely recapitulating prior academic history rather than detecting early-warning signals from current behaviour.

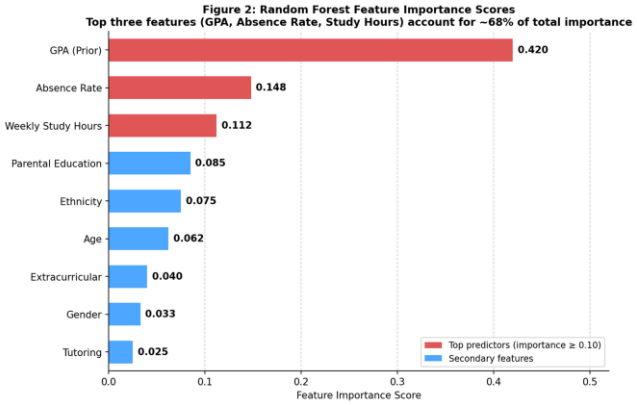

Fig 2: Random Forest feature importance scores. Prior GPA dominates at 0.420, followed by Absence Rate (0.148) and Weekly Study Hours (0.112). The top three features account for approximately 68% of total importance mass.

4.4 Per-Class Performance

Table 3 presents per-class Precision, Recall, and F1-Score for all models.

Table 3. Per-class Precision, Recall, and F1-Score for all models (n = 479)

Model	Grade	Prec.	Rec.	F1	Sup.
Rand. Forest	Grade A	0.56	0.57	0.57	60
	Grade B	0.41	0.42	0.42	95
	Grade C	0.41	0.38	0.39	93
	Grade D/F	0.51	0.46	0.48	74
Log. Regression	Grade A	0.64	0.75	0.69	60
	Grade B	0.60	0.59	0.60	95
	Grade C	0.58	0.55	0.57	95
	Grade D/F	0.61	0.58	0.60	74
Grad. Boosting	Grade A	0.67	0.77	0.72	60
	Grade B	0.62	0.60	0.61	95
	Grade C	0.59	0.58	0.59	95
	Grade D/F	0.62	0.59	0.61	74

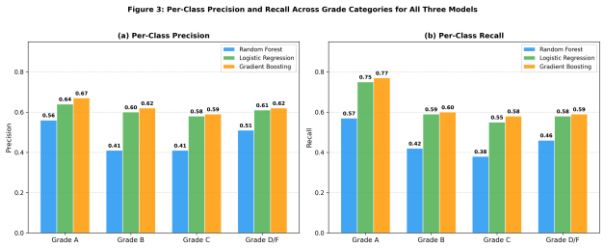

Fig 3: Per-class Precision (a) and Recall (b) for all three models across the four grade categories. Models perform best at the distribution extremes (Grade A and Grade D/F) and show lower performance for intermediate grades (B and C).

4.5 Confusion Matrix Analysis

Confusion matrix analysis reveals that most misclassifications occur between adjacent grade categories, particularly between grades B and C, and C and D/F. Logistic Regression demonstrated more balanced predictions across classes, while Gradient Boosting achieved higher precision but misclassified some minority-class instances. Random Forest exhibited the highest misclassification rates, especially for lower-performing students.

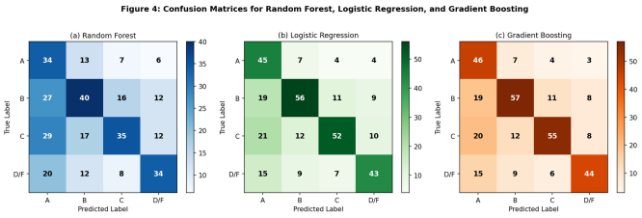

Fig 4: Confusion matrices for Random Forest (a), Logistic Regression (b), and Gradient Boosting (c). Diagonal entries represent correct classifications; off-diagonal concentrations along adjacent cells indicate systematic confusion between neighbouring grade categories.

5. DISCUSSION

5.1 Interpretation of Moderate Accuracy

The aggregate accuracy values (0.40–0.51) are notably lower than some prior work where accuracies of 0.80 and above are reported [3]. However, a random classifier on this four-class problem would achieve approximately 25% accuracy, making the best model's 0.51 approximately twice the random baseline. The multiclass nature of the task (four grade categories) is intrinsically more difficult than the binary classification (pass/fail) that dominates prior literature. These results are presented as an honest baseline rather than a claim of strong predictive performance.

5.2 Why Random Forest Underperforms

Random Forest achieves strong performance in high-dimensional feature spaces by building diverse decision trees on random feature subsets. In low-dimensional settings (15 features in this study), feature randomisation produces less diverse trees, reducing the ensemble's ability to average out errors. Gradient Boosting avoids this problem by building trees sequentially, each correcting residual errors of the last. Logistic Regression, as a linear model, benefits from the normalised feature matrix and class-weight balancing, performing comparably to Gradient Boosting despite lower model complexity — suggesting the signal in the current dataset is largely linearly separable.

5.3 Limitations

Dataset origin: The Kaggle dataset is synthetic and may not reflect the distributional properties of real institutional student records. Feature scope: The absence of behavioural trace data constrains the predictive ceiling of all models. Hyperparameter tuning: Models were evaluated at default or lightly configured settings; systematic grid search may improve performance, particularly for Random Forest. Static features: All features are static snapshots rather than time-series data. Single dataset: Validation on a single dataset limits generalisability.

6. PROPOSED EXTENSION: MULTIMODAL CNN INTEGRATION

6.1 Motivation

The results reveal a performance plateau at approximately 0.51 accuracy under structured-data-only conditions. Handwritten assignment submissions contain visual signals — spatial organisation, annotation density, revision marks, and handwriting regularity — that may correlate with cognitive engagement and academic performance in ways not captured by attendance or grade records. Incorporating these signals through CNN feature extraction represents a theoretically motivated and technically feasible extension to the current framework.

6.2 Architecture

The proposed extension follows a multimodal fusion architecture in which two feature streams are concatenated before a final classification layer. The first stream processes structured tabular features through a fully connected layer. The second stream processes scanned images of handwritten assignment submissions through a pre-trained CNN (ResNet-50 with ImageNet initialisation), fine-tuned on educational document images. Features from both streams are concatenated at the penultimate layer and passed through a softmax classification head trained end-to-end on the grade prediction task.

6.3 Expected Contributions

(1) A novel multimodal dataset combining structured student records with paired handwritten assignment images, collected with institutional ethical approval. (2) Empirical evaluation of CNN feature extraction strategies (fine-tuned ResNet-50 vs. vision transformer embeddings) for educational document images. (3) Ablation analysis comparing the contribution of the visual stream to structured-data-only and multimodal fusion models. (4) A reproducible open-source implementation enabling adoption by the learning analytics research community.

7. CONCLUSION

This paper presented a systematic comparative evaluation of Random Forest, Logistic Regression, and Gradient Boosting for student academic performance prediction under a unified preprocessing framework. Gradient Boosting achieved the best overall accuracy (0.51) and precision (0.51), while Logistic Regression demonstrated the highest recall (0.52); Random Forest underperformed both at 0.40 accuracy. A structured analysis identified dataset size, feature diversity, and the absence of visual or behavioural data as the primary constraints on predictive performance.

These findings establish a transparent, reproducible baseline and directly motivate a proposed multimodal extension incorporating CNN-extracted features from handwritten assignment images — a substantive and novel research direction with the potential to break through the accuracy plateau demonstrated here.

8. ACKNOWLEDGMENTS

The authors thank the Department of Computer Science, Saginaw Valley State University, for support during this research. The student performance dataset used in this study is publicly available via Kaggle [7].

9. REFERENCES

[1] S. D. A. Bujang et al., "Multiclass Prediction Model for Student Grade Prediction Using Machine Learning," IEEE Access, vol. 9, pp. 95608-95621, 2021.

[2] P. Asthana et al., "Prediction of Student's Performance with Learning Coefficients Using Regression Based Machine Learning Models," IEEE Access, vol. 11, pp. 72732-72742, 2023.

[3] R. Al-Shabandar et al., "Detecting At-Risk Students With Early Interventions Using Machine Learning Techniques," IEEE Access, vol. 7, pp. 149464-149478, 2019.

[4] Y. Zhang, X. Peng, and T. Huang, "Multimodal Learning Analytics: Combining Trace and Artefact Data for Student Outcome Prediction," Computers & Education, vol. 184, p. 104520, 2022.

[5] F. Alzubaidi et al., "Towards Unified Deep Learning for Multimodal Medical Data," NPJ Digital Medicine, vol. 6, no. 1, pp. 1-13, 2023.

[6] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD, 2016, pp. 785-794.

[7] W. A. Qayyum, "Student Performance Dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/waqi786/student-performance-dataset