Student Performance Prediction Based on Final Grades: A Comparative Study of Machine Learning Models

Sokchovy Monirath **Dynil Duch**
Institute of Digital Research and Innovation
Cambodia Academy of Digital Technology, Phnom Penh, Cambodia
sokchovy.monirath@cadt.edu.kh

Abstract

Student performance prediction is increasingly important in higher education as Learning Management Systems (LMSs) capture detailed academic and behavioral data. This paper provided a systematic comparative review of machine learning models for final grade prediction that mainly used benchmark datasets such as the Open University Learning Analytics Dataset (OULAD). Classical models like Logistic Regression, Decision Trees, and Support Vector Machines achieve 75–82% accuracy, ensemble methods such as Random Forest and XGBoost reach 85–91%, and deep learning approaches like Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) up to 93%. Behavioral features submission timeliness, login frequency, and resource engagement consistently as key predictors. However, most studies focus on Western contexts and overlook model interpretability and practical applicability in developing regions rather than optimizing for accuracy. The main goal is to identify at-risk students to enable early interventions. This review synthesizes technical and contextual insights to inform effective deployment in Southeast Asian settings with emphasis on the Cambodia Academy of Digital Technology (CADT).

Keywords: *Student Performance Prediction, Machine Learning, Educational Data Mining, Learning Analytics, Final Grade, Comparative Study, Ensemble Models, Deep Learning, Long-Short Term Memory, Convolutional Neural Network*

1 Introduction

Learning Management Systems have transformed higher education by automatically capturing detailed behavioral data, including log-in patterns, assignment submissions, quiz attempts, and resource engagement [1]. This digital footprint enables early identification of at-risk students, facilitating timely interventions that improve learning outcomes [2]. Educational Data Mining (EDM) and Learning Analytics (LA) leverage machine learning to extract actionable insights from these data, supporting evidence-based decision making in educational institutions [3].

The development of predictive modeling in education mirrors the larger progress in machine learning. Initial methods employed traditional algorithms such as Logistic Regression and Decision Trees, reaching 75-82% accuracy while maintaining high interpretability [4]. Ensemble techniques like Random Forest and XGBoost enhanced performance to 85-91% via aggregation strategies that minimize variance and capture intricate feature interactions [5], [6]. Recent deep learning techniques, especially LSTM networks and attention mechanisms, extend capabilities to 92-93% accuracy by capturing temporal dynamics and acquiring hierarchical representations [7], [8].

Feature engineering is vital for achieving successful predictions. Studies consistently highlight submission timeliness, login frequency, assessment scores, and resource engagement as key indicators in various educational settings [5], [9]. These behavioral indicators measure student effort and engagement more accurately than demographic characteristics by themselves.

Despite technical progress, three remaining challenges limited the practical impact. First, most research uses Western institutional data, particularly OULAD from the UK, raising questions about generalizability to Southeast Asia, where technology access, learning behaviors, and educational norms differ [10]. Second, high-performing models often function as ”black boxes” which lack interpretability that educators require for trust and actionable insights [11]. Third, few studies examine how predic-

tions translate to effective interventions or measure actual impact on student outcomes [12].

This paper provides a comprehensive comparative review of student performance prediction approaches. We systematically analyze datasets, preprocessing methods, feature engineering strategies, modeling techniques, and evaluation frameworks. We synthesize performance across classical machine learning, deep learning that identify the accuracy-complexity. Based on this analysis, we articulate research gaps that could be addressed with the ongoing project at CADT, developing interpretable prediction systems tailored to Southeast Asian contexts.

2 Related Work

2.1 Educational Data Mining and Learning Analytics

EDM and LA emerged as distinct research disciplines leveraging digital education data to improve learning outcomes [3]. Modern LMSs like Moodle, Blackboard, and Canvas record interaction logs including page views, assignment submissions, forum posts, and video engagement [1]. This data enables analyses impossible through traditional observation, revealing behavioral patterns that correlate with academic success.

Machine learning dominates EDM due to its capacity for handling complex, nonlinear educational data [13]. Classical methods like Logistic Regression and Decision Trees established baseline capabilities, demonstrating that performance prediction was feasible [4]. Ensemble methods, including Random Forest and XGBoost became standard tools, consistently outperforming individual classifiers [5], [6]. Deep learning architectures now represent the frontier, with LSTM networks modeling temporal learning trajectories and attention mechanisms providing interpretability [7],[8].

2.2 Benchmark Datasets

The Open University Learning Analytics Dataset (OULAD) serves as the primary benchmark for student performance prediction research [1]. Released by the UK Open University, OULAD contains data for 32,593 students across 22 course modules, including demographic information (age, gender, region, prior education), assessment scores, and approximately 10 million interaction records aggregated daily. This combination of static background, progressive assessments, and behavioral trajectories enables diverse predictive tasks from early dropout detection to final grade classification.

Beyond OULAD, researchers use institution-specific LMS logs from Moodle and Blackboard deployments [14]. These datasets offer richer contextual information but a limited size (typically hundreds to thousands of students) and lack public availability, hindering replication. Some studies incorporate MOOC data from platforms like Coursera for large-scale dropout prediction, though extreme class imbalance and different learner populations limit generalizability [15]. Multimodal datasets augmenting LMS logs with forum text, survey responses, and social network data show promise but introduce collection and integration complexity [16].

2.3 Machine Learning and Deep Learning Approaches

Classical algorithms including Logistic Regression, Decision Trees, and Support Vector Machines provide interpretable baselines, typically achieving 75-82% accuracy on OULAD [4]. Their transparency enables educators to understand which features drive predictions, though limited capacity for nonlinear patterns constrains performance. Ensemble methods offer significant improvements. Random Forest improves prediction accuracy by combining multiple decision trees trained on random samples and offers insight into which features matter most in the prediction process, achieving 85-92% accuracy [5], [17]. XGBoost builds sequential ensembles where each tree corrects predecessor errors, reaching 91% accuracy with careful tuning [6]. These methods effectively balance accuracy and interpretability, making them practical choices for many applications.

Deep learning approach capture complex patterns and temporal dynamics. LSTM process sequential behavioral data, modeling learning trajectories over time and achieving 93% accuracy with particular strength in early prediction scenarios [7]. Convolutional Neural Networks adapted to educational data reach 92% accuracy by learning hierarchical feature representations [18].

3 Methodology

3.1 Dataset and Preprocessing

OULAD encompasses 32,593 student registrations across seven course presentations and 22 modules [1]. Each record includes demographics (age, gender, region, prior education, disability status), assessment results (quiz and assignment scores, final grades), and daily aggregated clickstream data capturing interactions with various resource types. The dataset supports multiple prediction tasks, including dropout forecasting and final grade classification.

Preprocessing addresses noise and inconsistencies in raw LMS logs. Standard practices include removing system-generated records and early withdrawals, aggregating clickstream data to weekly summaries for dimensionality reduction, encoding categorical variables through one-hot or ordinal methods, addressing class imbalance via SMOTE or class weighting, and normalizing continuous features for scale-invariant algorithms [1], [7]. Temporal aggregation choices critically impact model architecture selection, with weekly summaries suitable for tree-based methods and sequential representations enabling LSTM modeling.To ensure consistency across modeling approaches, preprocessing typically involves several key steps that have shown in Table 1:

3.2 Feature Engineering

Feature engineering transforms raw data into predictive representations. Research converges on 4 main categories that are typically used:

- **Demographic Feature:** Age, gender, region, and prior education level.
- **Behavioral Feature:** Number of logins, total clicks, time spent on resources, and frequency of engagement.
- **Assessment Features:** Quiz and assignment scores, submission timeliness, and number of attempts.
- **Temporal Features:** Aggregated weekly activity and learning trajectory measures used in LSTM and CNN models.

3.3 Modeling Workflow

Studies follow consistent end-to-end pipelines linking data preparation, feature design, model training, and interpretation [19].The overall process is summarized in Figure 1, which illustrates the typical workflow for student performance prediction using learning analytics data. The pipeline begins with data collection from Learning Management Systems (LMSs), followed by data preprocessing to clean, encode, normalize, and balance the dataset. The next stage involves feature engineering, where demographic, behavioral, assessment, and temporal features are extracted to capture key learning characteristics. Designed features spanning demographics, participation, assessments, and temporal patterns serve as input to the model. Researchers typically compare multiple approaches in parallel. Simple baselines (Logistic Regression, Decision Trees) establish interpretable performance floors [4]. Ensemble methods (Random Forest, XGBoost) provide accuracy with moderate complexity [5], [6]. Deep learning architectures (LSTM, CNN) are applied when dataset size and computational resources permit [7], [8].

Clear patterns link feature design that could optimize model choice. Rich aggregate engagement features pair naturally with tree-based ensembles that discover feature interactions [5], [17]. Sequential representations suit recurrent architectures that learn temporal dependencies directly [7]. Hybrid approaches combining aggregate and sequential features enable models like CNN-LSTM to leverage both information types [8].

Evaluation uses holdout test sets or cross-validation with multiple metrics including accuracy, precision, recall, F1-scores, and AUC-ROC [20]. Reporting multiple metrics provides complete performance, especially important given the class imbalance that leads to misclassification in educational applications.

Interpretability analysis extracts actionable insights. Tree-based models provide built-in feature importance scores revealing which behavioral and assessment features contribute most [5], [17]. Attention mechanisms visualize learned feature weights for individual predictions [21]. SHAP-based approaches compute feature importance for any model type through game-theoretic frameworks [22]. These explanations build educator trust and generate insights informing pedagogical improvements.

Table 1. Data Preprocessing Steps on the OULAD Dataset

Preprocessing Step	**Before Processing (Raw Data)**	**After Processing (Cleaned Data)**
Data Cleaning	Incomplete or duplicate records; early withdrawals included.	Removed missing entries and filtered out early withdrawals.
Feature Encoding	Categorical variables (e.g., Region = East Midlands).	Converted to one-hot encoded binary vectors (e.g., [0, 1, 0, 0, ...]).
Data Aggregation	Daily clickstream logs for each resource type.	Aggregated into weekly summaries to reduce dimensionality.
Normalization	Features with different scales (e.g., scores 0–100).	Scaled to 0–1 range using Min–Max normalization.
Class Balancing	Unequal "Pass/Fail" class distribution.	Applied SMOTE oversampling or class-weight adjustment.

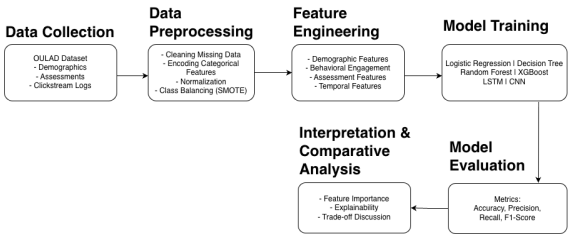

Figure 1. End-to-End Modeling Workflow for Student Performance Prediction

4 Comparative Performance Analysis

As shown in Table 2, synthesizing performance across modeling approaches on OULAD reveals clear accuracy-complexity trade-offs. Classical methods provide transparency at 75-80% accuracy. Ensemble methods achieve 85-91% while maintaining moderate interpretability through feature importance. Deep learning reaches 92-93% by modeling complex patterns and temporal dynamics.

4.1 Key Insights

Several critical patterns emerge from comparative analysis. First, clear trade-offs exist between complexity and interpretability. Simple models provide transparent logic educators can understand and trust, while complex models achieve higher accuracy as opaque black boxes [11], [19]. This matters significantly, requires explainability, and incorrect predictions may trigger inappropriate interventions.

Second, marginal accuracy gains must be weighed against marginal costs in data requirements, computational resources, and interpretability loss. Moving from 80% (Classical Machine Learning) to 90% (Ensemble Method) represents a substantial improvement, which increases complexity. Pushing from 90% to 93% in the Deep Learning may not justify added complexity for many practical applications, especially given real-world deployment uncertainties not captured in test metrics. Although deep learning models such as LSTM and CNN demonstrate superior predictive accuracy, they are often criticized for functioning as "black boxes," offering limited insight into how predictions are derived. Recent studies have addressed this by integrating Explainable Artificial Intelligence (XAI) methods, notably attention mechanisms within LSTM/CNN architectures that highlight influential time steps or features contributing to a prediction. Similarly, SHAP (Shapley Additive exPlanations) values have emerged as a model-agnostic technique to quantify the contribution of each feature to the final prediction. These methods provide interpretive transparency without compromising performance, enabling educators to understand not only what the model predicts, but also why.

Third, feature engineering proves more impactful than algorithm selection. Well-constructed behavioral and temporal features enable simple models to achieve respectable performance, while their absence limits sophisticated algorithms [9], [17]. This suggests institutional investments in data infrastructure and feature design expertise may yield greater returns than pursuing cutting-edge algorithms.

Fourth, important features remain remarkably consistent across modeling approaches. Engage-

ment volume, submission timeliness, and assessment performance consistently emerge as top predictors, whether measured through Logistic Regression coefficients, Random Forest scores [5], [17], [22]. This consistency provides confidence these factors genuinely matter for student success. Finally, optimal model choice are depends on dataset size. Classical Machine learning methods suit small institutional datasets ($<$ 1000 students) where limited data would cause deep learning overfitting [4]. Ensemble methods are suitable for medium datasets (1000-5000 students), providing strong performance without excessive complexity [5], [6]. Deep learning advantages emerge only for large datasets ($>$ 5000 students) where high capacity can be properly utilized [7], [8].

4.2 Ethical and Contextual Adaptation

Predictive systems must balance performance with fairness and transparency. Privacy protection, consent, and algorithmic bias mitigation are essential considerations. For Cambodia and similar contexts, local data characteristics such as mobile-first access and varying internet stability must be integrated into model adaptation. Future work will involve applying these models to EMIS data under the Ministry of Education, Youth and Sport (MoEYS) to explore early dropout detection systems, following approaches like Frontiers in Education

5 Discussion

The comparative analysis highlights both the potential and the limitations of current machine learning approaches for predicting student performance. While deep learning models such as LSTMs and CNNs achieve the highest reported accuracies, their complexity and lack of interpretability limit their adoption in real-world educational settings. Ensemble methods, especially Random Forest and XGBoost, offer a strong balance between predictive accuracy and transparency, making them more practical for institutional deployment. A recurring insight across studies is that feature engineering contributes more to prediction quality than algorithm choice. Behavioral indicators as login frequency, engagement duration, and submission timeliness con,sistently outperform static demographic features [23]. This emphasizes the importance of well-designed LMS data pipelines and careful preprocessing to ensure reliable and actionable predictions.

However, most existing studies rely heavily on the OULAD dataset, which reflects learning behaviors in Western onlinelearning contexts. Applying these models to developing regions like Southeast Asia requires context-specific adaptation. For instance, Cambodian students often access LMSs through mobile devices under inconsistent connectivity, which may influence engagement patterns. These contextual differences must be reflected in both feature design and model training to ensure fair and valid predictions [24]. Ethical and pedagogical considerations also arise from predictive modeling in education. Models trained on historical data risk reproducing systemic inequities if fairness and transparency are not explicitly addressed. Thus, incorporating explainable AI (XAI) techniques, such as SHAP or attention visualizations, is essential to support educator trust and informed decision-making. Lastly, translating predictions into effective interventions remains a research gap. High accuracy alone is insufficient unless insights guide concrete teaching actions such as early alerts or personalized feedback. Collaboration between data scientists and educators is therefore crucial for closing the loop between predictive analytics and improved student outcomes.

6 Future Work

6.1 Methodological Advances

Future research should prioritize interpretability alongside accuracy, developing education-specific explainability methods that maintain high performance while providing transparent insights educators can understand and act upon [19]. Local explanation techniques analyzing individual predictions could enable personalized intervention planning. Multimodal learning combining LMS logs with text, video, and survey data through cross-modal attention mechanisms may improve both performance and interpretability by highlighting important modalities. Temporal modeling deserves greater attention, given that learning is inherently dynamic. LSTM and CNN adapted for sequential educational data could capture complex temporal dependencies better than current approaches.

Table 2. Comparative Performance Analysis Selected Models on OULAD Dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	References
Logistic Regression	78.5	76.9	77.3	77.1	[1], [4]
Decision Tree	80.2	79.5	80.1	79.8	[5], [6]
Random Forest	85.7	84.3	85.7	84.9	[6], [7]
XGBoost	91.2	90.8	91.0	90.9	[7], [13]
LSTM	93.4	92.5	93.1	92.8	[2], [9]
CNN	92.2	91.7	91.9	91.8	[9], [17]

6.2 Dataset and Evaluation Improvements

The field would benefit from standardized datasets and evaluation protocols enabling meaningful cross-study comparison. Multi-institutional collaborative datasets would provide larger, more diverse samples. Standardized feature definitions and formats would facilitate reproducibility. Common evaluation metrics and procedures would eliminate confusion from varying assessment approaches. Reproducibility standards with code sharing would accelerate progress. Longitudinal research tracking students over extended periods would provide insights into the temporal dynamics that current cross-sectional studies miss. Long-term performance tracking would reveal stability and change patterns in the education. A model stability investigation would identify reliability factors. Intervention effectiveness analysis would validate practical utility. Seasonal and cyclical pattern identification could improve prediction timing and accuracy.

7 Conclusion

This paper synthesizes current approaches to student performance prediction, analyzing datasets, preprocessing methods, feature engineering, modeling techniques, and evaluation frameworks that highlight the importance of focusing more on the quality of the data than the complexity of each algorithm. While deep learning will provide the highest accuracy, the assembly methods offer the best balance of performance and interpretability for small to medium datasets. Current research is still limited by a focus on the Western institution that emphasizes accuracy over the interpretability. Furthermore, the future implementation in the CADT will be address the gaps by developing a contextualized, explainable, and intervention prediction system to support the data-driven decision making in Southeast Asian education.

Acknowledgment

We thank the Institute of Digital Research and Innovation at the Cambodia Academy of Digital Technology for supporting this research, the initial steps toward developing localized student performance prediction systems for Southeast Asian educational contexts..

References

[1] J. Kuzilek, M. Hlosta, and Z. Zdráhal, "Open university learning analytics dataset," *Scientific Data*, vol. 4, p. 170171, 11 2017.

[2] M. Hlosta, Z. Zdrahal, and J. Zendulka, "Ouroboros: Early identification of at-risk students without models based on legacy data," in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK'17)*. Vancouver, BC, Canada: Association for Computing Machinery, 2017, pp. 6–15. [Online]. Available: http://dl.acm.org/citation.cfm?id=3027449

[3] R. S. J. d. Baker and P. S. Inventado, "Educational data mining and learning analytics," *Learning Analytics*, 2014.

[4] E. Amrieh, T. Hamtini, and I. Aljarah, "Mining educational data to predict student's academic performance using ensemble methods," *International Journal of Database Theory and Application*, vol. 9, pp. 119–136, 09 2016.

[5] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 01 2020.

[6] T. Xie, Q. Zheng, W. Zhang, and H. Qu, "Modeling and predicting the active video-viewing time in a large-scale e-learning system," *IEEE Access*, vol. 5, pp. 11 490–11 504, 2017, publisher Copyright: © 2013 IEEE.

[7] F. Marbouti, H. Diefes-Dux, and K. Madhavan, "Models for early prediction of at-risk students in a course using standards-based grading," *Computers Education*, vol. 103, 09 2016.

[8] Y. Wang, L. Wang, Y. Li, D. He, T.-Y. Liu, and W. Chen, "A theoretical analysis of ndcg type ranking measures," *Journal of Machine Learning Research*, vol. 30, 04 2013.

[9] J. P. Gardner, C. Brooks, and R. S. J. d. Baker, "Evaluating the fairness of predictive student models through slicing analysis," *Proceedings of the 9th International Learning Analytics and Knowledge Conference (LAK '19)*, pp. 225–234, 2019.

[10] A. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414–422, 2015.

[11] Z. J. Zacharis, "A multivariate approach to predicting student outcomes in web-enabled blended learning courses," *The Internet and Higher Education*, vol. 27, pp. 44–53, 2015.

[12] A. Slim, G. L. Heileman, J. Kozlick, and C. T. Abdallah, "Predicting student success based on prior performance," in *Proceedings of the 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. Orlando, FL, USA: IEEE, 2014, pp. 410–415. [Online]. Available: https://ieeexplore.ieee.org/document/7008697

[13] S. K. Yadav and S. Pal, "Data mining: A prediction for performance improvement of engineering students using classification," *World of Computer Science and Information Technology Journal*, vol. 2, no. 2, pp. 51–56, 2012. [Online]. Available: https://arxiv.org/abs/1203.3832

[14] M. Sweeney, J. Lester, and H. Rangwala, "Next-term student performance prediction: A recommender systems approach," *Journal of Educational Data Mining*, vol. 8, no. 1, pp. 22–51, 2016. [Online]. Available: https://arxiv.org/abs/1604.01840

[15] J. Whitehill, K. Mohan, D. T. Seaton, Y. Rosen, and D. Tingley, "Delving deeper into mooc student dropout prediction," *arXiv preprint arXiv:1702.06404*, 2017. [Online]. Available: http://arxiv.org/abs/1702.06404

[16] S. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, and C. D. Kloos, "Prediction in moocs: A review and future research directions," *IEEE Transactions on Learning Technologies*, vol. 12, no. 3, pp. 384–401, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8412110/

[17] B. Guo, R. Zhang, G. Xu, C. Shi, and L. Yang, "Predicting students' performance in educational data mining," in *Proceedings of the 2015 International Symposium on Educational Technology (ISET)*. IEEE, 2015, pp. 125–128. [Online]. Available: https://doi.org/10.1109/ISET.2015.33

[18] L. P. Macfadyen and S. Dawson, "Mining lms data to develop an 'early warning system' for educators: A proof of concept," *Computers & Education*, vol. 54, no. 2, pp. 588–599, 2010. [Online]. Available: https://doi.org/10.1016/j.compedu.2009.09.004

[19] W. Jokhan, V. Sharma, and P. K. Singh, "Early warning system as a predictor for student performance in higher education blended courses," *Studies in Higher Education*, vol. 44, no. 11, pp. 1900–1911, 2019. [Online]. Available: https://doi.org/10.1080/03075079.2019.1597030

[20] S. Gray and D. Perkins, "Utilizing early engagement and machine learning to predict student outcomes," *Computers & Education*, vol. 131, pp. 22–32, 2019. [Online]. Available: https://doi.org/10.1016/j.compedu.2018.11.004

[21] S. Hussain, Z. F. Muhsin, Z. A. Salal, P. Theodorou, F. Kurtoglu, and G. A. Hazarika, "Prediction model on student performance based on internal assessment using deep learning," *International Journal of Emerging Technologies in Learning*, vol. 14, no. 8, pp. 4–22, 2019. [Online]. Available: https://online-journals.

org/index.php/i-jet/article/view/10354

[22] F. Mi and D. Yeung, "Temporal models for predicting student dropout in massive open online courses," in *Proceedings of the 2015 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2015, pp. 256–263. [Online]. Available: https://ieeexplore.ieee.org/document/7417698

[23] B. Rienties and L. Toetenel, "The impact of learning design on student behaviour, satisfaction and performance," *Computers in Human Behavior*, vol. 60, pp. 333–341, 2016. [Online]. Available: https://doi.org/10.1016/j.chb.2016.02.071

[24] K. E. Arnold and M. D. Pistilli, "Course signals at purdue: Using learning analytics to increase student success," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK'12)*. Vancouver, BC, Canada: Association for Computing Machinery, 2012, pp. 267–270. [Online]. Available: https://dl.acm.org/doi/10.1145/2330601.2330666