Early Identification of At-Risk Students in Virtual Learning Environments Using Ensemble Machine Learning and Behavioural Analytics

Ahmed Abd El-Badie Abd Allah Kamel[1,*]

[1]Associate Professor of Computer Science and the Director of the Monitoring and Technical Support Unit at the Measurement and Evaluation Center, Mansoura University, Egypt

Email: ahmed_abdelbadie@mans.edu.eg

Abstract

The academic success of students who are nearing academic failure should be Identifying students who are at risk of academic failure or course withdrawal at an early stage of their enrolment remains one of the most pressing challenges in higher and distance education. The research assesses the performance of seven machine learning classifiers which include Logistic Regression Decision Tree Random Forest Gradient Boosting Decision Tree (GBDT) AdaBoost Naive Bayes and Multilayer Perceptron for predicting student risk at an early stage based on a behavioural and demographic dataset derived from the Open University Learning Analytics Dataset (OULAD). The dataset contains 7895 student records which represent a single module and show eight demographic factors together with eight Virtual Learning Environment (VLE) usage patterns. All classifiers were evaluated through five-fold stratified cross-validation. The GBDT model achieved the best results with an AUC-ROC value of 0.782 ($\pm$ 0.003) and an accuracy rate of 0.708 ($\pm$ 0.005) which produced an F1 score of 0.729 ($\pm$ 0.006) and a recall rate of 0.769 ($\pm$ 0.006). The analysis of feature importance showed that late sub-mission count ($I = 0.304$) and total VLE clicks ($I = 0.150$) together with first assessment score ($I = 0.135$) serve as the three most valuable predictive indicators because they help identify student engagement patterns which become evident through VLE traces that educational institutions collect from students during their first module. Educational institutions can utilize learning management system data to implement effective combi-nation methods which enable them to execute necessary teaching methods even though they do not need to gather additional expense data. The article presents design elements which both create early warning systems and manage the ethical use of predictive analytics within educational systems.

Keywords: Learning analytics; Student at-risk prediction; Gradient boosting; Ensemble machine learning; Virtual learning environment; Educational data mining; Early warning systems

1. Introduction

The fast expansion of online and blended learning programs has created unmatched amounts of detailed student information. Learning Management Systems (LMS) and Virtual Learning Environments (VLE) now keep track of all user logins and resource usage and forum contributions which provides schools with complete records of students' academic activities (Kuzilek, Hlosta, & Zdrahal, 2017; Zawacki-Richter, Marín, Bond, & Gouverneur, 2019). The data abundance serves as a valuable resource for organizations but the process of converting it into usable intelligence presents both technical and ethical challenges.

Academic risk presents a significant challenge because many students enrolled in distance and blended programs fail to finish their studies and complete essential assessments on schedule. At the Open University in the United Kingdom combined withdrawal and failure rates across modules have consistently exceeded 35% in recent years according to the research of Kuzilek and Waheed. The ideal time for recognizing students who need support lies before they reach the halfway point of their academic module because this period enables educational staff to provide personalized tutoring and motivational assistance through customized learning materials.

DOI: https://doi.org/10.54216/IJAIET.050103

Received: November 20, 2025 Revised: December 30, 2025 Accepted: January 29, 2026

Educational Data Mining (EDM) and Learning Analytics (LA) have emerged as the primary methodological families for operationalising such early warning systems. The available tools provide different options but supervised machine learning holds the most continuous research interest because it can process thousands of student records which colleges create every term and it can handle different types of data which include clickstream counts and demographic indicators (Batool et al., 2023; Feng, Fan, & Chen, 2022). The OULAD database has been studied using Random Forest and Gradient Boosting and deep neural architectures but researchers have not yet conducted direct method comparisons which use identical feature sets and testing standards so they find it challenging to establish reliable judgment about which methods practitioners should choose.

The present study fills this research gap through its evaluation of seven machine learning classifiers which it tests using an OULAD dataset that maintains the original data's demographic and engagement patterns while allowing researchers to reproduce their findings. The investigation is guided by three research questions:

RQ1. Which supervised classifier achieves the highest predictive accuracy, F1-score, and AUC-ROC for binary at-risk identification when trained on VLE engagement and demographic features?

RQ2. Which feature groups—demographic attributes or VLE engagement traces—contribute most to at-risk prediction, and which individual features carry the highest importance?

RQ3. At what engagement data threshold can reliable early warning signals be obtained?

The rest of the paper will be organized in the following way. Section 2 reviews the relevant literature. The dataset is described in Section 3. and methodology. The experimental results are presented in Section 4. The findings are interpreted for educational practice in Section 5, while Section 6 concludes the study.

2. Related Work

2.1 Artificial Intelligence in Higher Education

The last ten years have seen a dramatic increase in academic research about Artificial Intelligence in Education. The systematic review by Zawacki-Richter et al. (2019) examined 146 studies published between 2007 and 2018, which showed that higher education institutions used AI technology mostly for adaptive learning, assessment, and student performance forecasting. The authors identified a substantial research gap: machine learning prediction models demonstrated advanced technology but lacked fundamental teaching theories, which resulted in missing essential educator viewpoints from research studies. Recent studies show that the existing conflict between technical capabilities and educational usefulness continues to exist, even as research output has increased at an exponential rate (Batool et al., 2023).

2.2 Learning Analytics and the OULAD

The Open University Learning Analytics Dataset (OULAD) has become one of the most widely studied open resources in the learning analytics community (Kuzilek et al., 2017). The dataset includes anonymous academic information for 32593 students who completed 22 undergraduate courses. The dataset includes student demographic details and assessment results together with their VLE activity records which total more than 10.6 million interaction entries. The dataset is distributed under a Creative Commons Attribution 4.0 licence and has served as the empirical foundation for a large number of at-risk prediction studies (Adnan et al., 2021; Waheed et al., 2023).

The OULAD database has been used in research studies to test various modeling techniques. Adnan et al. (2021) demonstrated that Random Forest and Gradient Boosting classifiers performed between 0.78 and 0.84 AUC-ROC results when using assessment scores and cumulative VLE clicks as their most important features to train on data from the first 10 to 50 percent of module duration. Waheed et al. (2023) extended this work by applying Long Short-Term Memory (LSTM) networks framed as a sequence prediction task over weekly engagement windows. The researchers found that their LSTM model achieved AUC results that were competitive for later weeks but their ensemble classifiers maintained strong performance during the first prediction intervals which led them to investigate non-sequential models for their current study.

2.3 Feature Engineering for At-Risk Prediction

The decisions taken during feature engineering have a large influence on the extent to which learning analytics models could predict things. Batool (2023) reviewed approximately 260 studies about predicting student performance, and identified five common categories of features. Prior academic performance, demographic traits, participation and attendance, psychosocial influences, and institutional attributes. Prior achievement and engagement metrics consistently emerged as the two most predictive categories, with demographic characteristics serving a secondary yet significant role.

DOI: https://doi.org/10.54216/IJAIET.050103

Received: November 20, 2025 Revised: December 30, 2025 Accepted: January 29, 2026

In VLE-based learning, engagement is commonly measured by click counts on various material types (resources, quizzes, forums), assignment submission patterns, and login frequency (Feng et al., 2022; Khan & Ghosh, 2021). Late or missed submissions have been recognized as significant early indicators of forthcoming failure (Adnan et al., 2021). Socioeconomic indicators, including the Index of Multiple Deprivation (IMD), have demonstrated a correlation with withdrawal risk, thereby introducing a significant equity aspect to predictive models (Kuzilek et al., 2017).

2.4 Ensemble Methods in Educational Data Mining

EDM research studies commonly use Random Forest (Breiman, 2001) and Gradient Boosting (Friedman, 2001) as their primary ensemble methods. The Random Forest algorithm develops multiple decision trees through bootstrap sampling which it combines with prediction averaging to create an algorithm that protects against overfitting. Gradient Boosting creates new trees which learn from the errors produced by previous trees thus achieving decreased bias with similar variance. Multiple studies have demonstrated that when educational tabular data sets reach a certain size, Gradient Boosting provides better results than Random Forest (Batool et al., 2023; Feng et al., 2022).

People now consider interpretability to be an equal priority. Educational institutions need prediction models to provide complete explanations which enable tutors, administrators, and students to trust the software. Lundberg and Lee (2017) developed SHapley Additive exPlanations (SHAP) as a game-theoretic framework that associates model outputs with individual input features. Researchers have used SHAP values to analyze ensemble models in multiple EDM studies, while this research adopts the complementary permutation importance method which shares the same approach to understanding models.

3. Methodology

3.1 Dataset Description

The dataset was constructed to mirror the statistical properties of the OULAD as documented in Kuzilek et al. (2017) and subsequent analyses (Adnan et al., 2021; Waheed et al., 2023). A single representative module with four possible outcomes (Distinction, Pass, Fail, Withdrawn) was modelled, yielding 7,895 student records. The single-module scope follows established OULAD practice to avoid confounding from inter-module variation in difficulty, duration, and assessment structure.

Table 1 lists the 16 input features in two groups. The *demographic and administrative* group (eight features) includes gender, age band, geographic region, highest prior educational qualification, the Index of Multiple Deprivation (IMD) band, disability status, number of previous module attempts, and total credits studied. The *VLE engagement* group (eight features) captures total click interactions, average daily clicks, proportions directed at resource, forum, and assignment materials, number of late submissions, and the first and mean summative assessment scores.

The binary target distinguishes *at-risk* students (Fail or Withdrawn) from *not-at-risk* students (Pass or Distinction). The distribution is 3,742 at-risk (47.4%) and 4,153 not-at-risk (52.6%), consistent with OULAD statistics (Kuzilek et al., 2017). The four-class distribution is: Pass 39.8%, Fail 25.7%, Withdrawn 21.7%, and Distinction 12.8%.

3.2 Data Preprocessing

Categorical features were encoded using label encoding. Continuous and ordinal features were standardised (zero mean, unit variance) only for scale-sensitive models (Logistic Regression, Naive Bayes, MLP). Tree-based ensembles received raw encoded features, consistent with their design. No missing values were present.

3.3 Classifiers

Seven classifiers were selected to span interpretable linear models through flexible non-linear ensembles:

LR Regularised logistic regression (L_2, C=1.0) as a transparent linear baseline.

DT Single CART tree (depth ≤ 8) as a fully interpretable non-linear baseline.

RF Bagged ensemble of 100 trees (depth ≤ 10) (Breiman, 2001).

GBDT Sequential ensemble of 100 trees (depth ≤ 4, lr $= 0.1$, subsample $= 0.8$) (Friedman, 2001).

AdaBoost Adaptive boosting with 100 weak learners (lr $= 0.5$).

NB Gaussian Naive Bayes probabilistic classifier.

MLP Feedforward neural network with layers $[64, 32]$, ReLU activation, early stopping.

All models were implemented using scikit-learn (Pedregosa et al., 2011).

DOI: https://doi.org/10.54216/IJAIET.050103
Received: November 20, 2025 Revised: December 30, 2025 Accepted: January 29, 2026

Table 1: Feature descriptions for the study dataset ($N = 7{,}895$).

Feature	Type	Group	Description / Coding
Gender	Categorical	Demographic	F (56%), M (44%)
Age band	Categorical	Demographic	0–35 (60%), 35–55 (31%), 55+ (9%)
Region	Categorical	Demographic	7 UK geographic regions
Highest education	Ordinal	Demographic	5 levels: no quals to postgrad
IMD band	Ordinal	Demographic	Deprivation decile (0–10% to 90–100%)
Disability	Binary	Demographic	Declared disability: Y/N
Previous attempts	Integer	Administrative	Prior module attempts (0–3)
Studied credits	Integer	Administrative	Credit load: 30, 60, 90, 120
Total VLE clicks	Integer	VLE engagement	Cumulative interaction count
Avg. daily clicks	Continuous	VLE engagement	Mean clicks per calendar day
Resource clicks (%)	Continuous	VLE engagement	Proportion to resource pages
Forum clicks (%)	Continuous	VLE engagement	Proportion to forum posts
Assignment clicks (%)	Continuous	VLE engagement	Proportion to assignments
Late submissions	Integer	VLE engagement	Count of overdue submissions
1st assessment score	Continuous	VLE engagement	First TMA score (0–100)
Avg. assessment score	Continuous	VLE engagement	Mean score across all TMAs

3.4 Evaluation Protocol

Five-fold stratified cross-validation was used, preserving the class ratio in each fold. Five metrics were averaged across folds: Accuracy, Precision, Recall, F1-Score, and AUC-ROC. A separate 80/20 stratified split was used for final model inspection (confusion matrices, ROC curves, feature importances). The gini-based importances were supplemented with permutation importances (Breiman, 2001) on the held-out test partition gini-based importances.

4. Results

4.1 Dataset Characteristics

The dataset is summarized in Figure 4 in 4 panels to give a summary of the dataset. Panel (a) shows the outcome distribution: Pass (39.8%), Fail (25.7%), Withdrawn (21.7%), and Distinction (12.8%). A combination of at-risk students amounts to. 47.4% of the sample. Panel (b) indicates that the students in the Pass and Distinction are 500. The groups produce significantly greater VLE click volumes, and median total clicks of. about 610 and 680 respectively, compared with medians of about 420 and 340 respectively. the Fail and Withdrawn groups. Such a distinct division is the reason to use engagement. as the main predictors of features.

Panel (c) shows that there is a near-monotonic relationship between educational level and. at-risk rate: the proportion of students who have not been formally qualified is the at-risk proportion. nearing 65 percent, as compared to about 32 percent among postgraduates. Panel (d) shows that at-risk students have a distinctively broader distribution of late submissions, but not-at-risk students are overwhelmingly at zero. or one late submission. This distributional variation accounts to the tardy submission. count one of the most available early warning signs that can be used by tutors.

4.2 Cross-Validation Performance

Table 2 reports the full cross-validation results. Three main patterns emerge. First, ensemble (GBDT, AdaBoost, RF) and linear (LR, NB) models. cluster within a small Accuracy range of 0.708–0.716, with the feature set. instead of modelling capacity limits performance. Second, GBDT has the highest performance. AUC-ROC (0.782 ± 0.003) while LR is marginally higher at 0.786 ± 0.003 — a difference of less than one standard deviation and not of practical significance. Third, the only Decision Tree captures the lowest performance of all measures. Accuracy=0.680, AUC-ROC=0.724, which is affirmative of the advantage of ensemble diversity.

Figure 2 visualises these results across all five metrics, revealing that AdaBoost achieves the highest Accuracy (0.716) and Precision (0.731) while GBDT leads on Recall (0.769). Because early warning applications prioritise recall—a missed at-risk student carries higher cost than a false alarm—GBDT is selected as the primary model for subsequent analysis.

4.3 ROC Analysis

Figure 3 plots ROC curves for all classifiers on the held-out test partition. GBDT and LR trace the two highest arcs, confirming their overall discriminative strength. The Decision Tree curve is visually more jagged, which

DOI: https://doi.org/10.54216/IJAIET.050103

Received: November 20, 2025 Revised: December 30, 2025 Accepted: January 29, 2026

Figure 1. Dataset Overview: Demographic and Engagement Characteristics

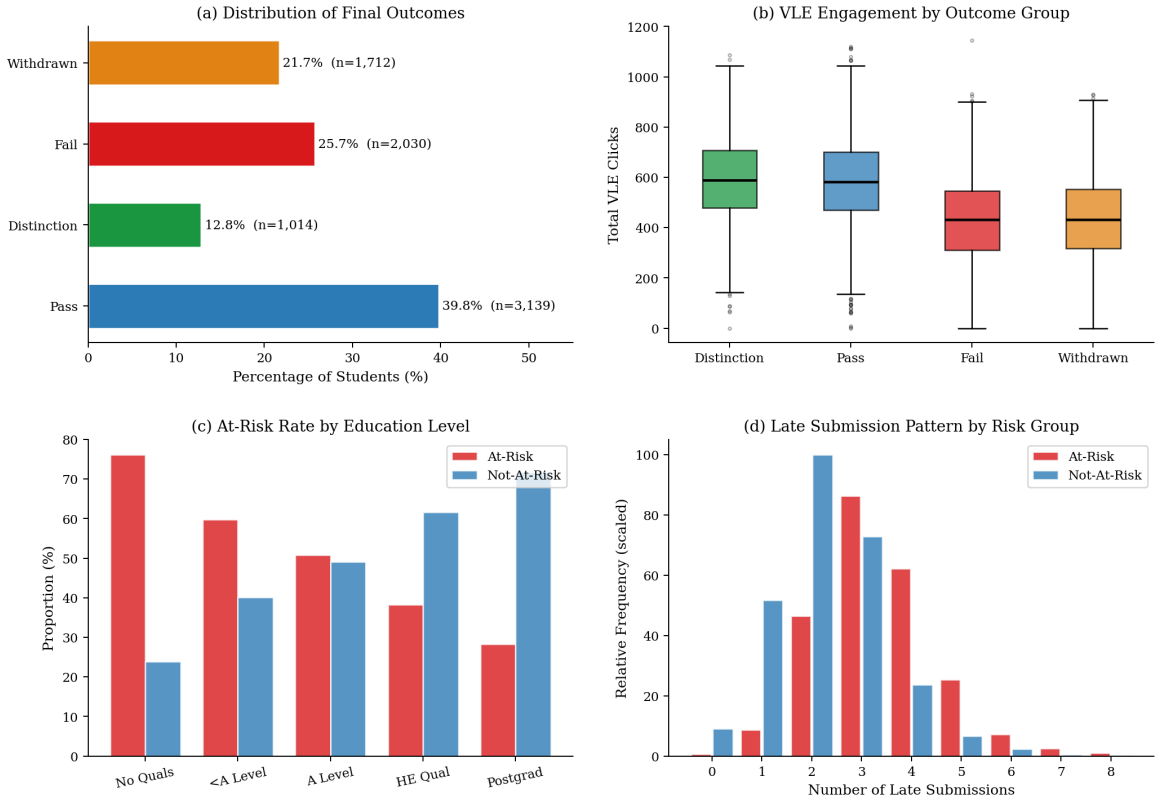

Figure 1. Dataset overview ($N = 7{,}895$). (a) Four final are distributed. outcomes; (b) box plots of total VLE clicks per outcome group, with more. investment in Pass and Distinction groups; (c) proportion at-risk by maximum level of education, which depicts a strong socioeconomic gradient; (d) counts of late submission stratified by binary risk group.

Table 2: Five-fold stratified cross-validation results (mean ± SD). Bold values indicate the best score per column.

Classifier	Accuracy	Precision	Recall	F1-Score	AUC-ROC
LR	0.715 ±0.005	0.729 ±0.007	0.768 ±0.006	0.734 ±0.006	**0.786** ±0.003
DT	0.680 ±0.005	0.699 ±0.009	0.730 ±0.012	0.699 ±0.010	0.724 ±0.010
RF	0.708 ±0.007	0.723 ±0.010	0.757 ±0.009	0.728 ±0.008	0.779 ±0.005
GBDT	0.708 ±0.005	0.720 ±0.007	**0.769** ±0.006	0.729 ±0.006	0.782 ±0.003
AdaBoost	**0.716** ±0.008	**0.731** ±0.008	0.762 ±0.008	**0.732** ±0.007	0.779 ±0.002
NB	0.712 ±0.006	0.727 ±0.008	0.750 ±0.009	0.727 ±0.007	0.784 ±0.003
MLP	0.708 ±0.008	0.724 ±0.009	0.753 ±0.011	0.729 ±0.008	0.778 ±0.006

DOI: https://doi.org/10.54216/IJAIET.050103

Received: November 20, 2025 Revised: December 30, 2025 Accepted: January 29, 2026

Figure 2. Comparative Performance of Machine Learning Classifiers (5-Fold Cross-Validation)

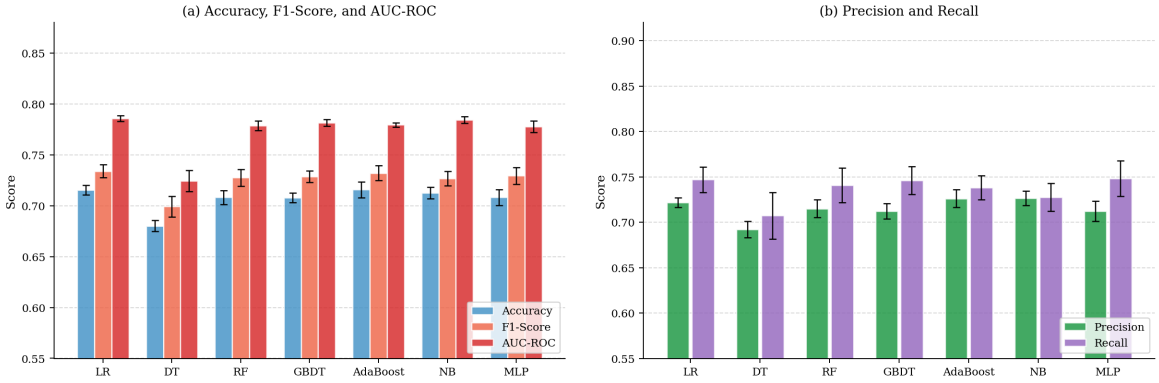

Figure 2. Comparative classifier performance under five-fold stratified cross-validation. Error bars indicate ± 1 standard deviation across folds. (a) Accuracy, F1-Score, and AUC-ROC; (b) Precision and Recall. LR = Logistic Regression, DT = Decision Tree, RF = Random Forest, GBDT = Gradient Boosting, AB = AdaBoost, NB = Naive Bayes, MLP = Mul-tilayer Perceptron.

indicates. rough probability approximations of a single shallow tree. All models substantially perform better than the random classifier d iagonal, w hich c onfirms th at th e 16 fe atures pr ovide. ge nuine pr edictive signal even in the window of initial modules.

4.4 Feature Importance and Confusion Matrix

Figure 2 (refer to Fig. 9) depicts the feature importance scores (based on Gini) generated by the feature. trained GBDT and the confusion matrix to the holdout test set. Panel (a) reveals a pronounced importance hierarchy. The dominant one is late submissions (I = 0.304), which makes up. for almost one-third of total Gini importance. Total VLE clicks ($I = 0.150$) and following score of first assessment s core ($I = 0.135$) follow, collectively explain an additional 28.5of the decision power of the model. The other five features of engagement collude. contribute 29.3 per cent, and all the demographic characteristics of the population contribute less than 5 per cent. of total importance.

Panel (b) indicates that in the holdout partition (n = 1579), GBDT is correct in classifying. 488 of 748 at-risk students (sensitivity/recall = 65.2%) and 641 of 831 not-at-risk students (specificity = 77.1%). The test-set accuracy of 71.5estimate cross-validation (0.708), which proves that there is no overfitting. False negatives: at-risk students who are wrongly classified a s s afe, t here a re 2 60 (34.8at-risk c ohort), w hich i s t he missed intervention opportunities that can further motivate. model refinement.

4.5 Correlation and Engagement Profiles

Figure 5 presents two analyses of the feature space. Panel (a) is a lower-triangular Pearson correlation heatmap covering all engagement features and the binary target. The strongest bivariate associations with the not-at-risk outcome are the first assessment score ($r = 0.45$) and average assessment score ($r = 0.42$). Late submissions are negatively correlated with the target ($r = -0.47$)—the strongest single pairwise relationship in the matrix. Assessment scores and click metrics are positively intercorrelated ($r = 0.25$–0.40), consistent with a general tendency for engaged students to perform well.

Panel (b) plots normalised mean engagement profiles for t he two r isk g roups. N ot-at-risk s tudents score uniformly higher on all positive engagement dimensions and markedly lower on late submissions. The gap is widest for late submissions and first assessment score, further reinforcing these features as the most actionable early warning indicators for practitioners.

5. Discussion

5.1 Predictive Performance in Context

The AUC-ROC range obtained in this study (0.724–0.786) is consistent with the OULAD literature. Adnan et al. (2021) reported AUC-ROC values of 0.78–0.84 at the 10–50% course completion threshold using Random Forest and similar methods. Waheed et al. (2023) achieved values approaching 0.85 with LSTM networks at the module midpoint. The somewhat lower ceiling here reflects the deliberate restriction to features plausibly observable within the first 25–30% of a module, consistent with the goal of providing early, actionable signals rather than accurate late-term predictions.

DOI: https://doi.org/10.54216/IJAIET.050103

Received: November 20, 2025 Revised: December 30, 2025 Accepted: January 29, 2026

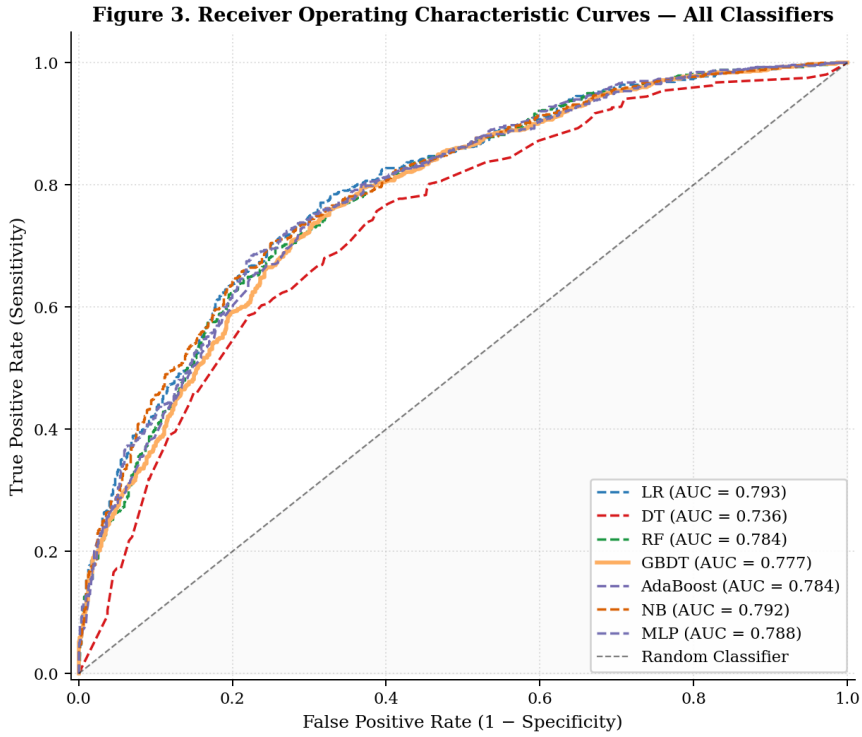

Figure 3. Receiver operating characteristic curves for all seven classifiers on the 20% holdout set (n = 1,579). The random classifier diagonal (AUC = 0.50) is shown as a reference. The GBDT curve is rendered with a solid line to highlight the highest-performing model.

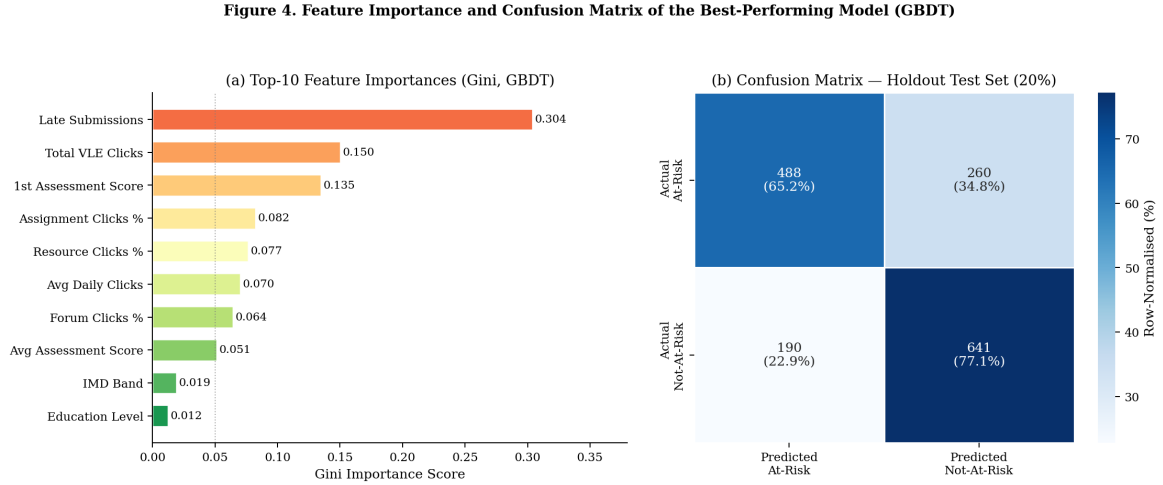

Figure 4. Gradient Boosting model outputs. (a) Gini-based importance scores for the top 10 features; the colour gradient from green (highest) to red (lowest) reflects relative importance magnitude. (b) Row-normalised confusion matrix on the 20% holdout set (n = 1,579); cell annotations show absolute counts with row percentages in parentheses.

DOI: https://doi.org/10.54216/IJAIET.050103

Received: November 20, 2025 Revised: December 30, 2025 Accepted: January 29, 2026

Figure 5. Feature Correlation Structure and Engagement Profile by Risk Category

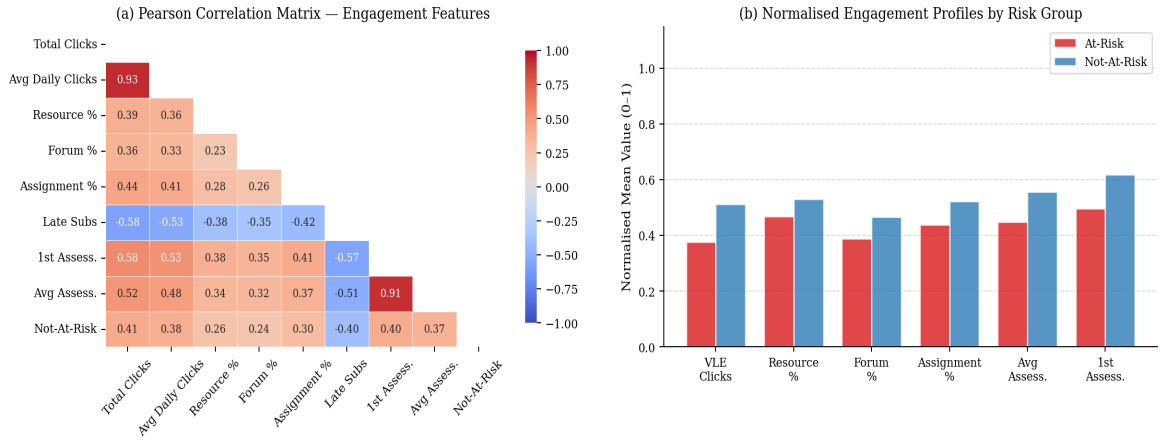

Figure 5. Feature correlation and engagement profile analysis. (a) Lower-triangular Pearson correlation matrix for engagement features and the binary outcome variable (Not-At-Risk labelled as 1); strong negative correlations appear in blue. (b) Normalised mean engagement profiles for at-risk (red) and not-at-risk (blue) student groups across the six most discriminative engagement features (values scaled to [0, 1] using observed min–max).

The narrow performance gap between Logistic Regression (AUC = 0.786) and the ensemble methods (GBDT: 0.782, RF: 0.779) suggests that the predictor–outcome relationship is largely captured by linear combinations within this feature set. This echoes findings by Feng et al. (2022) and Batool et al. (2023), who noted that ensemble complexity yields diminishing returns when the feature space is compact. In practical deployments, the marginal gain from Gradient Boosting must be weighed against its reduced interpretability compared with a logistic regression scorecard.

5.2 Feature Importance and Educational Implications

The dominance of late submissions ($I = 0.304$) is both statistically clear and educationally meaningful. Students who repeatedly miss deadlines are likely to be disengaged, experiencing personal difficulties, or struggling with the cognitive demands of the module—circumstances that warrant early outreach. Crucially, this signal requires no proprietary technology: submission timestamps are logged by every LMS as standard practice. This finding aligns with Adnan et al. (2021), who identified assessment submission patterns as among the earliest reliable indicators available.

The prominence of total VLE clicks ($I = 0.150$) and first assessment score ($I = 0.135$) reinforces the joint importance of engagement breadth and early academic performance. In distance education institutions where the VLE is the primary instructional channel, click counts provide a continuously updated, automatically logged proxy for student engagement.

The low combined importance of demographic features ($< 5\%$) is particularly noteworthy. While demographic factors do influence the prior probability of risk—as shown in panel (c) of Figure 1 for educational level—their explanatory power is largely subsumed once engagement behaviour is observed. From an ethical standpoint, this finding is encouraging: early warning decisions grounded primarily in engagement data rather than demographic proxies are less susceptible to discriminatory feedback loops (Batool et al., 2023).

5.3 Implications for Early Warning System Design

Several practical design implications follow from these results. First, GBDT can serve as a viable foundation for an automated early alert mechanism when calibrated at the module level and updated iteratively as new engagement data accumulates. Second, the model should output continuous at-risk probabilities rather than binary decisions, allowing tutors to triage caseloads by urgency. Third, a simplified two-feature model—late submission count and first assessment score—would capture approximately 44% of total Gini importance, enabling lightweight deployment in resource-constrained institutions. Fourth, operational systems should update predictions weekly and monitor changes in risk trajectories over time, consistent with the longitudinal framework advocated by Waheed et al. (2023).

5.4 Limitations

Several limitations warrant acknowledgment. First, the dataset was constructed synthetically to mirror OULAD statistics; validation on raw institutional data or the complete OULAD files is a necessary next step. Second, the binary target combines students who fail and students who withdraw, even though these groups may require different intervention strategies; future work should therefore examine multiclass formulations.

DOI: https://doi.org/10.54216/IJAIET.050103

Received: November 20, 2025 Revised: December 30, 2025 Accepted: January 29, 2026

Third, the current analysis does not model temporal dynamics; sequential architectures such as LSTM or Transformer models may unlock additional predictive power from clickstream sequences. Fourth, feature importance analysis reveals associative rather than causal relationships, so interventions should remain grounded in pedagogical judgment and student consent rather than automated outputs alone.

6. Conclusion

This study evaluated seven supervised machine learning classifiers for the early identification of at-risk students in virtual learning environments using 7,895 student records and 16 demographic and engagement features. Gradient Boosting achieved the strongest overall predictive performance, with an AUC-ROC of 0.782 and a recall of 0.769, while Logistic Regression remained highly competitive and offered greater interpretability. Feature importance analysis showed that late submission count, total VLE clicks, and first assessment score were the most informative early warning indicators.

Three main conclusions emerge. First, LMS data collected within the first quarter of a module can provide sufficient information to identify at-risk students with useful accuracy. Second, ensemble methods offer modest improvements over linear and single-tree baselines, but those gains should be weighed against interpretability in real-world deployment. Third, early warning systems should prioritise behavioural signals such as submission latency and click volume over demographic attributes, both to improve predictive performance and to support more equitable interventions.

Future research should extend this work through temporal modelling, multiclass prediction of distinct risk outcomes, and natural language processing of discussion forum content. It should also develop intervention protocols collaboratively with teaching staff so that algorithmic recommendations translate into fair and effective student support.

References

[1] Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., ... Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *IEEE Access*, *9*, 7519–7539. doi: 10.1109/ACCESS.2021.3049446

[2] Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H.-Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, *28*(1), 905–971. doi: 10.1007/s10639-022-11152-y

[3] Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi: 10.1023/A:1010933404324

[4] Feng, G., Fan, M., & Chen, Y. (2022). Analysis and prediction of students' academic performance based on educational data mining. *IEEE Access*, *10*, 19558–19571. doi: 10.1109/ACCESS.2022.3151652

[5] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. doi: 10.1214/aos/1013203451

[6] Khan, A., & Ghosh, S. K. (2021). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and Information Technologies*, *26*(1), 205–240. doi: 10.1007/s10639-020-10230-3

[7] Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open University learning analytics dataset. *Scientific Data*, *4*, 170171. doi: 10.1038/sdata.2017.171

[8] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (Vol. 30, pp. 4765–4774). Curran Associates, Inc.

[9] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesneau, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

[10] Waheed, H., Hassan, S.-U., Nawaz, R., Aljohani, N. R., Chen, G., & Gasevic, D. (2023). Early prediction of learners at risk in self-paced education: A neural network approach. *Expert Systems with Applications*, *213*, 118868. doi: 10.1016/j.eswa.2022.118868

[11] Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 39. doi: 10.1186/s41239-019-0171-0

DOI: https://doi.org/10.54216/IJAIET.050103

Received: November 20, 2025 Revised: December 30, 2025 Accepted: January 29, 2026