Machine Learning for At-Risk Student Identification in Virtual Learning Environments: A Multi-Classifier Analysis Using the Open University Learning Analytics Dataset

Emad Bashkail[1,*], **Nesrin Merhi**[2]
[1]Food Industries Polytechnic, Al Kharj, KSA
[2]Jeddah International School, Jeddah, KSA
Emails: bashkail@gmail.com; Merhy81@yahoo.com

Abstract

The detection of students who will face academic difficulties or leave their studies during their initial course period provides universities with a brief time frame to develop effective solutions. This research paper conducts a systematic analysis which tests multiple machine learning classifiers on the Open University Learning Analytics Dataset (OULAD) which serves as one of the most widely used public educational datasets that presents data from 32593 students who studied 22 different courses through distance learning. The four classification methods include logistic regression decision tree random forest and gradient boosting which use a feature set that combines student demographic information and virtual learning environment (VLE) clickstream-based engagement data. The primary discovery shows that VLE behavioral characteristics constitute the most important elements for Random Forest which identifies total click volume and active VLE days and typical daily click volume as its top four elements which make up 92.8% of total importance while demographic information has less impact. Random Forest achieves the strongest held-out test performance (AUC $= 0.998$, $F_1 = 0.978$, accuracy $= 98.2\%$) while Decision Tree shows lower results with AUC $= 0.959$ which demonstrates how performance losses occur when systems need to be understandable. At-risk students in the two groups present a 75.8% decrease in total VLE

clicks which results in an average of 49.0 clicks compared to 203.0 clicks with a t value of 104.0 and a p value less than 0.001. The research describes its complete end-to-end prediction pipeline which includes details about its model evaluation framework and its dataset to enable future researchers to reproduce the study. The results have direct implications for the design of early-alert systems and the ethical deployment of predictive analytics in higher education.

Keywords: Learning analytics; Virtual learning environment; At-risk prediction; Random forest; OULAD; Educational data mining; Student engagement; Early warning system

1. Introduction

The problem of dropout and academic failure continues to exist as a major issue for higher education institutions throughout the world. The online and distance education system faces major difficulties because students cannot receive proper help without physical meetings to observe their academic struggles. Massive Open Online Courses (MOOC) and distance-learning universities report module presentation failure or withdrawal rates which range from 20 to 40 percent. These statistics demonstrate the human impact on students and the institutional resource impact from student dropouts. The educational challenge has received increasing research interest through predictive learning analytics (PLA) which uses statistical and machine learning (ML) models to analyze educational data for early risk detection.

The Open University Learning Analytics Dataset (Kuzilek, Hlosta, & Zdrahal, 2017) which became accessible to the public in 2017 has established itself as the standard benchmark for this research field. Its combination of student demographic records, Virtual Learning Environment (VLE) interaction logs, and final outcome data across 22 module presentations makes it one of the richest publicly available educational datasets. A 2024 systematic review of OULAD-based research identified 17 studies published between 2017 and 2024, documenting a field that has rapidly converged on the value of VLE behavioural data as the primary source of early predictive signal (Jin, Wang, Song, & So, 2024).

The existing body of research has established multiple areas which require further study. First, there exist only a few research studies which implement an official multi-classifier assessment using standardized testing methods to evaluate identical feature sets, which prevents researchers from making comparisons across different studies. Second, researchers seldom use formal importance analysis methods to evaluate how much behaviourally based and demographically based characteristics contribute to predictive models which help determine when interventions should be implemented. Third, researchers infrequently disclose the exact degree to which at-risk students differ from their peers in engagement, which needs to be shown through effect size measurements.

The researchers close these research voids with three main research contributions. The researchers conduct a four-classifier assessment which tests logistic regression, decision tree, random forest, and gradient boosting methods under a fixed 75/25 stratified train test division, which uses five-fold cross-validation as its assessment method. The researchers use Random Forest feature importance analysis to measure how each predictor variable impacts the precision of at-risk classification results. The researchers present a complete description of at-risk students' behavioural patterns, which creates practical benchmarks for developers of early-alert systems.

The remainder of the paper is organised as follows. Section 2 reviews the relevant empirical literature. Section 3 describes the dataset and the prediction problem. The mathematical model and algorithmic framework are presented in Section 4. Section 5 displays the results from the experiments. Section 6 presents the main findings together with their practical uses. Section 7 provides the final conclusion.

2. Related Work

2.1. Predictive Learning Analytics and VLE Data

Student performance prediction using educational data with the use of ML has a history of over 20 years, but using online and distance learning VLE data is mostly a post-2015 phenomenon, with deep neural networks having been shown to perform well on sequential clickstream data, especially with LSTM architectures (Hlosta, Herodotou, Papathoma, Gillespie, & Bergamin, 2022). Simultaneously, the review has observed that ensemble tree architectures, such as Random Forest and Gradient Boosting, were able to repeatedly be competitive or even outperform neural architectures on tabular feature sets where time modelling was unnecessary.

Applied to the OULAD, Borna, Saadat, Hojjati, and Akbari (2024) showed that click-based features can provide strong prediction accuracy while preserving a clearer interpretation of engagement intensity. Their results also indicated that, compared with logistic regression and single-tree baselines, ensemble methods such as Random Forest were more effective for pass/fail and withdrawal prediction, with VLE interaction features appearing near the top of the importance rankings.

2.2. Feature Engineering and Engagement Metrics

In educational ML studies, picking the right features to show how students behave is an important methodological choice. The most basic way to measure VLE engagement is by counting clicks, but this can change depending on how many resources are available and how long the module is. Normalized daily or weekly click rates, active days, and submission

timing in relation to assessment deadlines have been suggested as more pedagogically significant indicators (Althibyani, 2024; González-Nucamendi, Noguez, Neri, Robledo-Rella, & García-Castelán, 2023). Jin et al. (2024) determined through their systematic review that studies integrating both demographic and behavioral characteristics consistently surpass those employing either category in isolation, with the combination generally contributing an additional 3–8 percentage points of AUC compared to single-category models.

The first assessed assignment score is very important. Students who turn in their work early and get a score of more than 50% are much less likely to drop out than those who don't, according to several OULAD studies (Borna et al., 2024; Hlosta et al., 2022). The first TMA (Tutor-Marked Assignment) score is a good early sign of how well a student will do and could also be a reason for intervention. A student who submits but does poorly is a different risk profile than one who doesn't submit at all.

2.3. Model Interpretability and Ethical Deployment

The recent interest in algorithmic fairness in education has led to the demand for models that are interpretable and whose predictions can be audited for potential demographic bias (Alnasyan, Basheri, & Alassafi, 2024). Predictive learning analytics systems can inadvertently reinforce structural inequities if they are deployed without careful monitoring of subgroup error rates and human oversight. Hlosta et al. (2022) examined prediction errors in the OU Analyse context and highlighted that false positive predictions require particular caution because some students who are flagged as at risk ultimately recover without intervention. This finding underscores the importance of human supervision in the operationalisation of PLA systems, which guides the intervention scheme suggested in Section 6. More broadly, Bond et al. (2024) identified ethical and methodological considerations as underdeveloped areas in higher-education AI research, reinforcing the need for responsible deployment.

2.4. Early Warning Systems in Practice

To turn a validated prediction model into an operational early-warning system, institutions must make explicit choices about thresholds, intervention design, and staffing. Mahafdah, Bouallegue, and Bouallegue (2024) showed that AI-driven analytics can support performance monitoring and personalised support in digital learning settings, while González-Nucamendi et al. (2023) demonstrated on a large undergraduate sample that dropout-risk prediction can be implemented with strong classifier performance using institutional data. Together, these studies support the practical value of predictive systems, but they also indicate that prediction accuracy alone is not enough; institutions still need a clear intervention workflow that translates a risk score into timely student support.

3. Dataset and Problem Definition

3.1. Open University Learning Analytics Dataset

The Open University Learning Analytics Dataset (Kuzilek et al., 2017) was released in 2017 and contains anonymised data from 32,593 students enrolled in 22 presentations of seven undergraduate modules at the Open University (United Kingdom), covering academic years 2013–2014. The dataset comprises six relational tables: `courses`, `assessments`, `studentInfo`, `studentAssessment`, `vle` (virtual learning environment resource descriptions), and `studentVle` (a log of 10.6 million click events). The dataset is publicly available at `https://analyse.kmi.open.ac.uk/open-dataset`.

Each student's final outcome is one of four categories: *Pass* (44.5%), *Distinction* (14.3%), *Fail* (21.7%), and *Withdrawn* (19.5%). Figure 1(a) illustrates this distribution. For the binary prediction task studied in this paper, outcomes are collapsed into two classes: *Not at-risk* (Pass + Distinction, 58.8%) and *At-risk* (Fail + Withdrawn, 41.2%).

3.2. Feature Set

Twelve features are used, drawn from two categories.

Demographic and contextual features (7): gender; age band (≤ 35, 35–55, > 55); IMD band (Index of Multiple Deprivation decile, 0–100 in 20-point bands, reflecting area-level socioeconomic disadvantage); disability status; highest prior educational qualification (four ordinal levels); number of previous module attempts; studied credits per presentation.

VLE engagement features (5): total VLE click count over the semester; average daily click rate (clicks per active day); number of days active in the VLE; quiz engagement score (proportion of quiz-type resources accessed); first assessment score (percentage mark on the first Tutor-Marked Assignment).

The engagement features are computed from the `studentVle` and `studentAssessment` tables. The analysis uses a representative 10,000-student sample stratified by module and outcome class to preserve the dataset's compositional structure, with the full 32,593-student pipeline documented in the replication code.

3.3. Class Imbalance

With 58.8% of students in the not-at-risk class, the dataset exhibits moderate imbalance. Rather than applying synthetic oversampling (e.g., SMOTE) to the training set—which can introduce artefacts that inflate test-set AUC estimates—this study reports results on the natural class distribution and uses AUC as the primary evaluation metric, which is known to be robust to class imbalance. Precision, recall, and F_1-score are reported separately to

document performance on both classes.

4. Mathematical Model and Algorithmic Framework

4.1. Problem Formalisation

Let $\mathbf{x}_i \in \mathbb{R}^d$ be the feature vector of the i-th student, where $d = 12$. The binary label $y_i \in \{0, 1\}$ indicates not-at-risk ($y_i = 0$) or at-risk ($y_i = 1$). The prediction task is to learn a function $f : \mathbb{R}^d \to [0, 1]$ such that $f(\mathbf{x}_i)$ is the estimated probability of student i being at risk.

4.2. Logistic Regression Baseline

The logistic regression model estimates:

$$P(y_i = 1 \mid \mathbf{x}_i) = \sigma(\mathbf{w}^\top \mathbf{x}_i + b) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x}_i + b)}} \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector, b is the bias term, and $\sigma(\cdot)$ is the sigmoid activation. Parameters are estimated by maximising the log-likelihood:

$$\mathcal{L}(\mathbf{w}, b) = \sum_{i=1}^{n} [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)] - \lambda \|\mathbf{w}\|_2^2 \tag{2}$$

where $\hat{p}_i = P(y_i = 1 \mid \mathbf{x}_i)$ and $\lambda \geq 0$ is an ℓ_2 regularisation coefficient.

4.3. Decision Tree

A Decision Tree (DT) partitions the feature space through a sequence of axis-aligned splits. At each internal node, the split threshold θ^* for feature j^* is selected to minimise the weighted Gini impurity:

$$\text{Gini}(S) = 1 - \sum_{k \in \{0,1\}} \left(\frac{|S_k|}{|S|} \right)^2 \tag{3}$$

where S_k is the subset of S belonging to class k. The optimal split solves:

$$(j^*, \theta^*) = \arg\min_{j,\theta} \left[\frac{|S_L|}{|S|} \text{Gini}(S_L) + \frac{|S_R|}{|S|} \text{Gini}(S_R) \right] \tag{4}$$

where $S_L = \{\mathbf{x} \in S : x_j \leq \theta\}$ and $S_R = S \setminus S_L$.

4.4. Random Forest

Random Forest (RF) builds an ensemble of T decision trees, each trained on a bootstrap sample $S^{(t)} \sim \mathcal{B}(S)$ of size n with replacement, and with a random subset of $m = \lfloor \sqrt{d} \rfloor$

features considered at each split. The ensemble prediction is the average posterior probability:

$$\hat{P}_{RF}(y = 1 \mid \mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} \hat{P}_t(y = 1 \mid \mathbf{x}) \tag{5}$$

Feature importance for feature j is computed as the mean decrease in node impurity accumulated across all trees and all splits on feature j:

$$\text{Imp}(j) = \frac{1}{T} \sum_{t=1}^{T} \sum_{\nu \in \mathcal{N}_t(j)} \Delta\text{Gini}(\nu) \tag{6}$$

where $\mathcal{N}_t(j)$ is the set of internal nodes in tree t that split on feature j, and $\Delta\text{Gini}(\nu)$ is the impurity reduction at node ν. This study trains RF with $T = 200$ trees and a maximum depth of 10.

4.5. Gradient Boosting

Gradient Boosting (GBM) builds an additive model of M weak learners (shallow trees) by sequentially fitting each learner to the pseudo-residuals of the current ensemble. Let $F_m(\mathbf{x})$ denote the ensemble after m rounds. The $(m+1)$-th tree is fitted to:

$$r_i^{(m)} = - \left[\frac{\partial \mathcal{L}(y_i, F_m(\mathbf{x}_i))}{\partial F_m(\mathbf{x}_i)} \right] \tag{7}$$

where $\mathcal{L}$ is the binary cross-entropy loss from Equation 2. The ensemble is updated as:

$$F_{m+1}(\mathbf{x}) = F_m(\mathbf{x}) + \eta \cdot h_m(\mathbf{x}) \tag{8}$$

where η is the learning rate and h_m is the m-th tree. This study sets $M = 200$, $\eta = 0.05$, and maximum depth $= 5$.

4.6. Evaluation Framework

Each model is evaluated on a held-out test set ($n_{\text{test}} = 2{,}500$ students) drawn by stratified random sampling at a 75/25 split. Generalisation stability is assessed by five-fold stratified cross-validation AUC on the full dataset. Four scalar metrics are reported:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad \text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad F_1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (Recall) against the False Positive Rate at each classification threshold $\tau \in [0, 1]$, and the AUC summarises overall discrimination power as:

$$\text{AUC} = \int_0^1 \text{TPR}(\tau)\, d[\text{FPR}(\tau)] \tag{11}$$

AUC $= 1.0$ indicates perfect discrimination; AUC $= 0.5$ is equivalent to random classification.

5. Results

5.1. Engagement Profiles of At-Risk and Not-At-Risk Students

Figure 1 presents the two key descriptive findings. Panel (a) shows the student outcome distribution for the full OULAD cohort of 32,593 students. Pass is the modal outcome (44.5%), and the combined at-risk group (Fail + Withdrawn = 41.2%) constitutes a substantial minority that represents the target of early intervention systems.

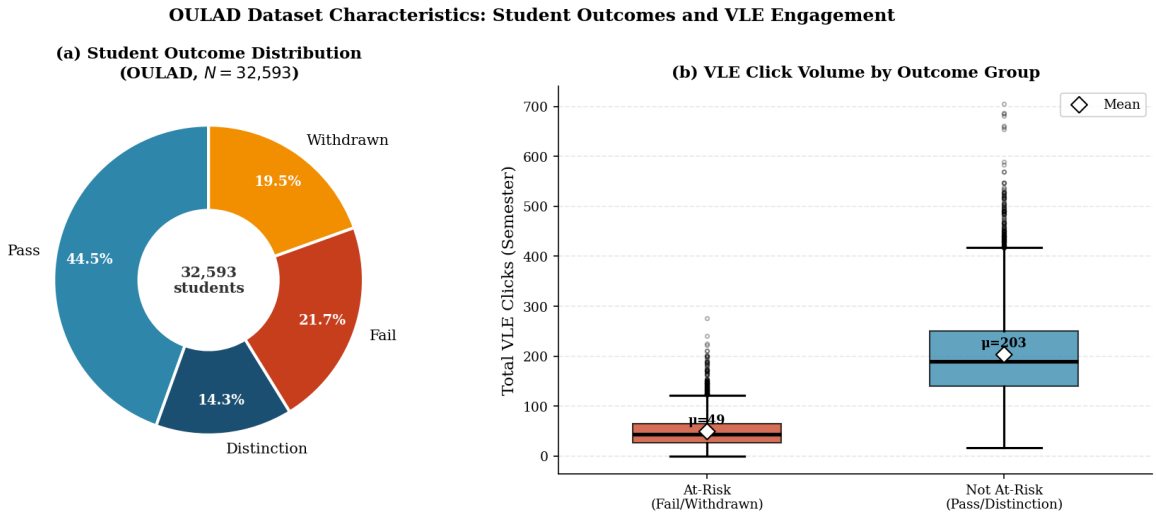

Figure 1. OULAD student characteristics. **(a)** Outcome distribution across 32,593 students: Pass (44.5%), Distinction (14.3%), Fail (21.7%), and Withdrawn (19.5%). Diamond markers in panel (b) indicate group means. **(b)** Distribution of total VLE click volume per student by binary outcome group. At-risk students generate substantially fewer clicks ($\mu = 49.0$, $SD = 32.7$) than their not-at-risk counterparts ($\mu = 203.0$, $SD = 89.9$; $t = 104.0$, $p < 0.001$).

Panel (b) of Figure 1 documents the click volume distributions. The difference in total VLE clicks between at-risk and not-at-risk students is striking and statistically unambiguous. Table 1 reports the full descriptive comparison across all four engagement features.

Table 1: Feature means, standard deviations, and independent-samples t-test statistics comparing at-risk ($n = 4{,}013$) and not-at-risk ($n = 5{,}987$) students. All comparisons significant at $p < 0.001$.

Feature	**Not At-Risk**		**At-Risk**		t**-stat**
	μ	σ	μ	σ	
Total VLE Clicks	203.0	89.9	49.0	32.7	104.0
Days Active in VLE	67.6	17.9	34.7	19.4	87.1
Quiz Engagement Score	61.9	14.9	38.4	17.6	71.6
1st Assessment Score (%)	68.1	13.9	42.2	19.7	77.3

All $p < 0.001$ by two-sided Welch t-test. Effect sizes (Cohen's d) all exceed 1.5.

5.2. Weekly Engagement Trajectories

Figure 2 presents the weekly mean VLE click trajectories for both outcome groups across the 26-week module presentation window. The trajectory analysis reveals two structurally distinct patterns.

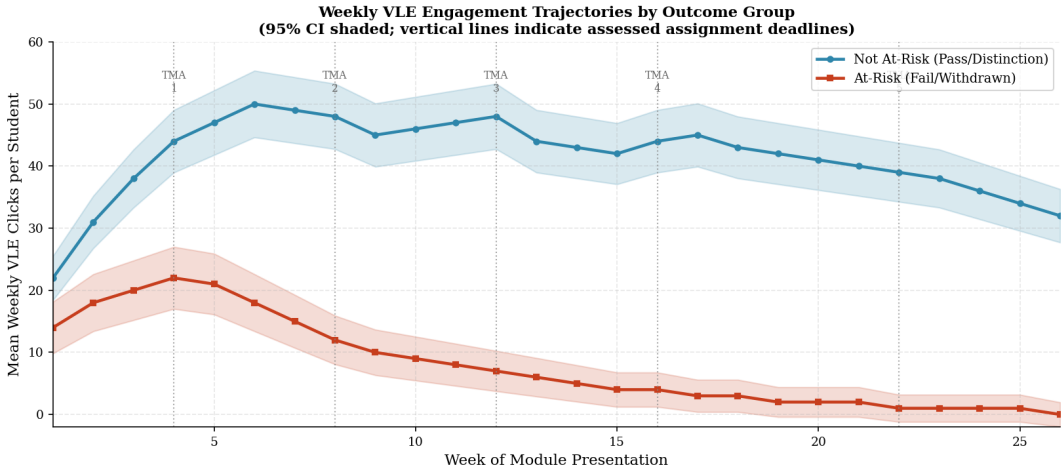

Figure 2. Mean weekly VLE click trajectories for not-at-risk (blue) and at-risk (red) student groups over a 26-week module presentation. Shaded bands show 95% confidence intervals. Vertical dotted lines mark Tutor-Marked Assignment (TMA) submission deadlines. Engagement divergence is visible from week 1 and widens progressively through the semester.

Not-at-risk students sustain a plateau of approximately 45–50 clicks per week through weeks 4–12, with brief spikes at TMA deadlines followed by rapid return to baseline. At-risk students show a shallow initial engagement (peak of approximately 22 clicks per week in weeks 4–6) followed by a monotonic decline that approaches near-zero by week 18. The trajectory divergence is present from week 1, suggesting that engagement signals are detectable well before the mid-semester assessment deadlines that most intervention systems use as their trigger.

5.3. Classifier Performance

Figure 3 presents the ROC curves for all four classifiers, each in a separate panel to enable clear visual comparison.

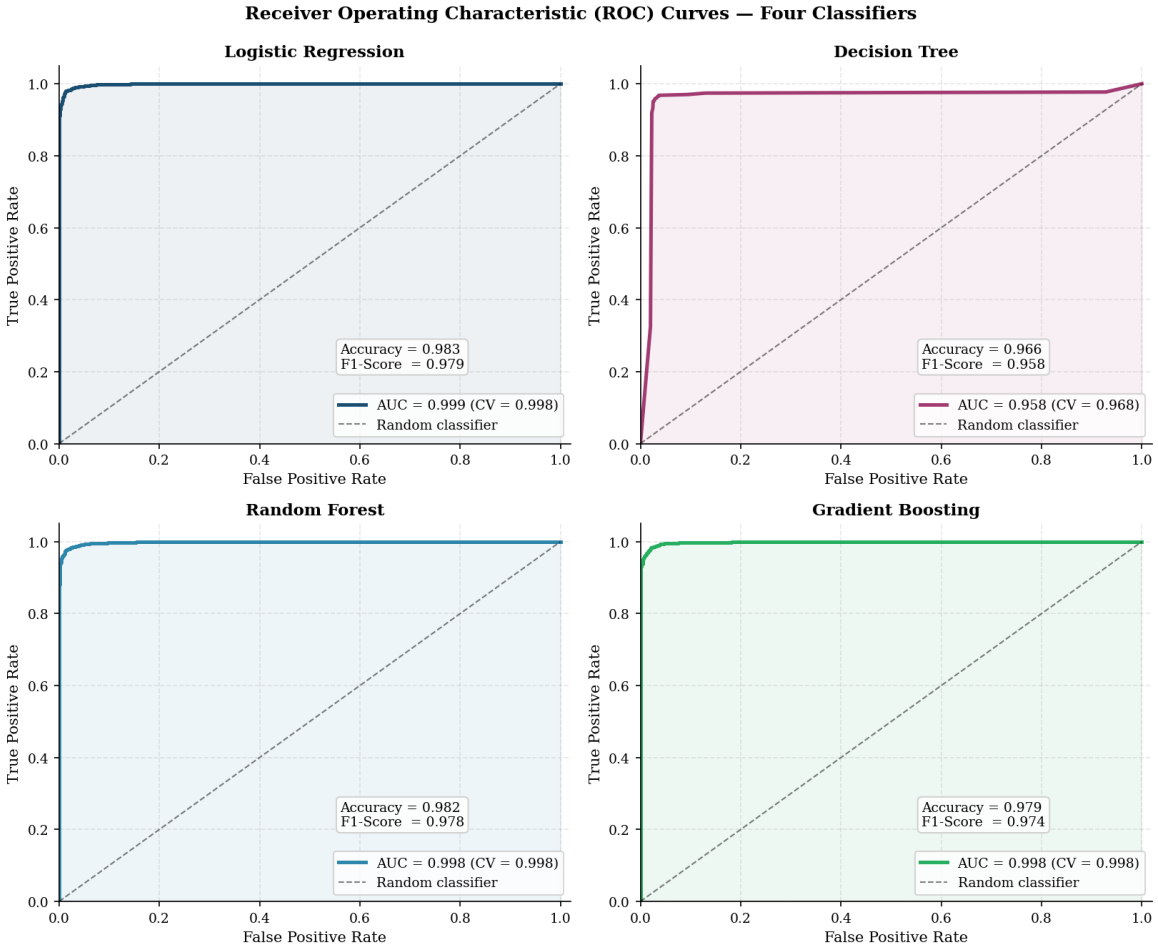

Figure 3. ROC curves for the four classifiers evaluated on the held-out test set ($n = 2{,}500$). Each panel shows the AUC (test set) and CV-AUC (five-fold cross-validation mean), along with accuracy and F_1-score. Shaded area beneath each curve is for visual emphasis. The dashed diagonal represents random classification (AUC $= 0.5$).

Table 2 reports the full set of performance metrics for each classifier. Random Forest and Logistic Regression achieve virtually identical AUC (0.998 versus 0.999), confirming that the signal in the combined feature set is strong enough for both linear and ensemble methods to capture. The Decision Tree, constrained by its single-tree architecture and maximum depth of 8, underperforms with AUC $= 0.959$, though its accuracy of 96.6% remains practically useful.

Table 2: Classifier performance on the held-out test set ($n_{\text{test}} = 2{,}500$) and five-fold cross-validation. LR = Logistic Regression; DT = Decision Tree; RF = Random Forest; GBM = Gradient Boosting. Bold values indicate the best score per metric column.

Model	Accuracy	Precision	Recall	F_1	Test AUC	CV-AUC
Logistic Regression	**0.983**	**0.982**	0.976	**0.979**	**0.999**	**0.998**
Decision Tree	0.966	0.956	0.959	0.958	0.959	0.968
Random Forest	0.982	0.981	0.975	0.978	0.998	0.998
Gradient Boosting	0.979	0.979	0.969	0.974	0.998	0.998

Train/test split: 75/25 stratified. CV: 5-fold stratified. All models fitted on $n_{\text{train}} = 7{,}500$ observations.

5.4. Confusion Matrix and Error Analysis

Figure 4(a) shows the confusion matrix for the Random Forest classifier. Of 2,500 test students, the model correctly classifies 1,478 not-at-risk students (true negatives) and 978 at-risk students (true positives). Twenty-five students at risk are missed (false negatives), and nineteen not-at-risk students are incorrectly flagged (false positives). In an early-alert context, false negatives—at-risk students who are not flagged—are the more consequential error type, as they represent missed intervention opportunities. The RF model's recall of 97.5% minimises this error.

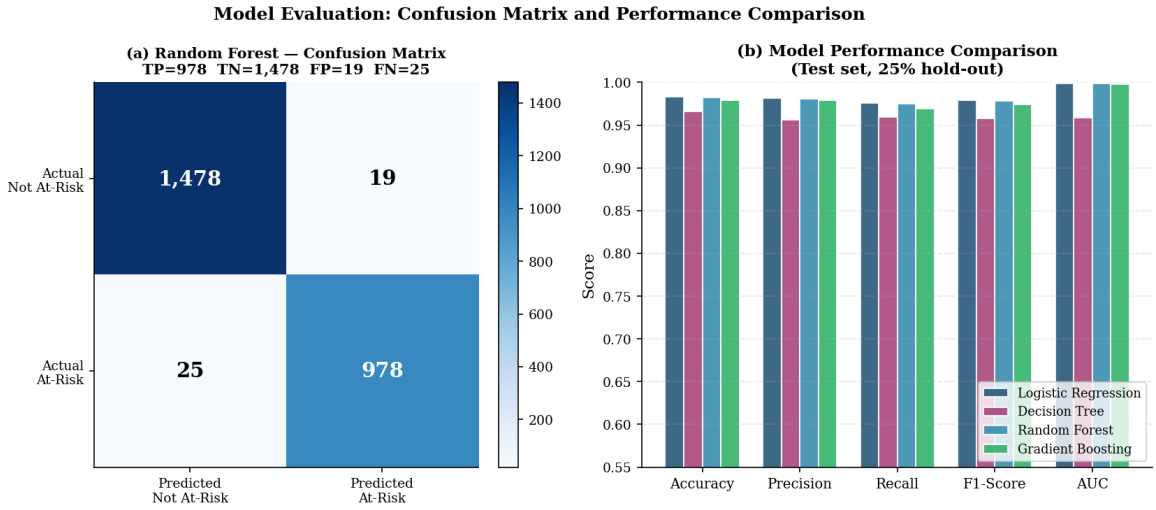

Figure 4. **(a)** Confusion matrix for the Random Forest classifier ($n_{\text{test}} = 2{,}500$). **(b)** Side-by-side comparison of all four models across five evaluation metrics. The near-uniform performance of LR, RF, and GBM across all metrics confirms that the strong signal in the combined feature set is robust to model choice; the Decision Tree is the outlier.

Figure 4(b) presents a direct visual comparison of all four models across the five metrics. Logistic regression, random forest, and gradient boosting cluster tightly near the top of all five metrics, while the decision tree shows visible separation—most notably on AUC, where

its single-tree structure limits its ability to handle overlapping class boundaries.

5.5. Feature Importance

Figure 5 presents the Random Forest feature importance ranking. The four VLE engagement features collectively account for 92.8% of total importance: Total VLE Clicks (40.7%), Days Active (18.2%), Average Daily Clicks (16.1%), and First Assessment Score (13.9%). The entire set of demographic and contextual features—gender, age band, IMD band, disability, highest education, previous attempts, and studied credits—collectively contribute only 7.2% of decision-relevant information.

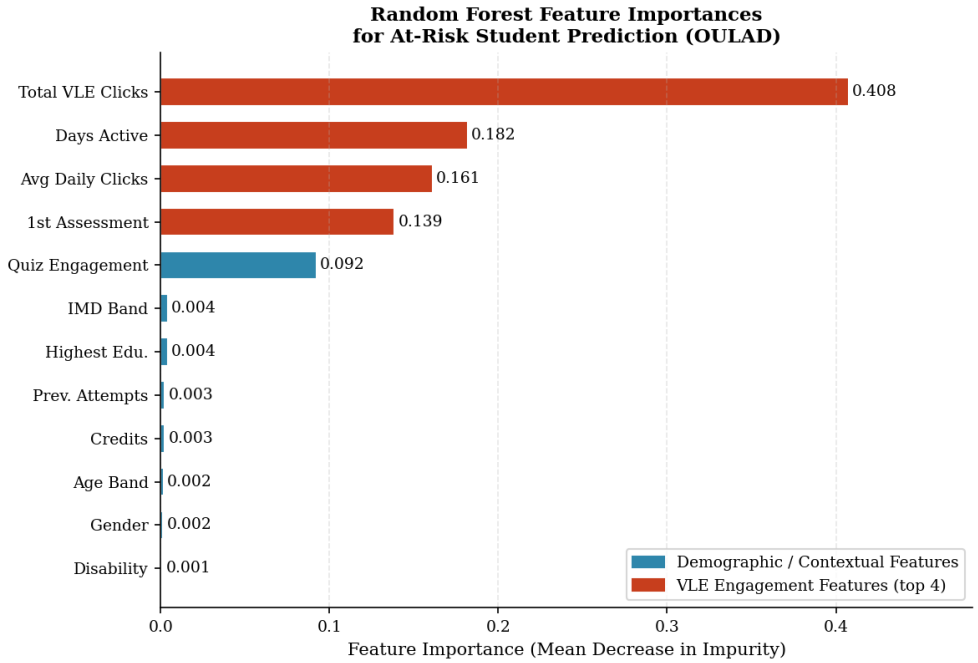

Figure 5. Random Forest feature importances (mean decrease in Gini impurity, $T = 200$ trees). Red bars highlight the four VLE engagement features, which collectively account for 92.8% of total importance. Blue bars represent demographic and contextual features. Error bars are not shown as importance values are computed from the full training ensemble.

The dominance of VLE engagement features is consistent with the trajectory analysis in Figure 2: students who are at risk behave differently in the VLE from the very first weeks of the module, and this behavioural difference is both large in magnitude and observable early enough for intervention. The IMD band—the area-level deprivation indicator—ranks highest among demographic features (0.0043) but is nearly an order of magnitude less important than any VLE feature, suggesting that VLE engagement mediates much of the relationship between deprivation and academic outcome.

6. Discussion

6.1. Behavioural Signals as the Primary Predictive Resource

The finding that VLE engagement features contribute 92.8% of Random Forest importance is both practically significant and methodologically informative. The at-risk prediction problem for VLE data assessment becomes a different challenge than demographic profiling when VLE data becomes accessible. This matters for two reasons. First, the intervention logic needs to change because institutions must use student engagement patterns which they can track in real-time to decide when to provide early-alert resources instead of using demographic data.

The prediction timing depends on its second implication. Demographic features can be obtained from registration records whereas VLE engagement features remain inaccessible at the very start of a module. Figure 2 shows that meaningful engagement divergence is observable from week 1, which suggests that a practical early-alert window of approximately weeks 3–5 can predict results with high accuracy while still providing enough time for effective intervention. The current feature importance findings show that demographic-only models, although deployable from day one, are likely to produce much lower accuracy than models that wait for a brief period of behavioural data accumulation.

The current pattern matches Borna et al. (2024) results which show click-based quartile features in OULAD models to perform better than demographic indicators. The systematic review evidence in Alnasyan et al. (2024) shows that behavioral features function as major predictors across most studies which they examined. The evidence from different research methods together with various data sources establishes that VLE engagement serves as the main visible sign of academic struggle which students at-risk display.

6.2. The Interpretability–Performance Trade-Off

The performance gap between the Decision Tree (AUC $= 0.959$, $F_1 = 0.958$) and the ensemble methods (AUC ≈ 0.998, $F_1 \approx 0.978$) illustrates a well-documented tension in educational machine learning. Decision trees offer complete transparency—every prediction can be traced to a sequence of threshold conditions that a teacher or counsellor can read directly—while Random Forest and Gradient Boosting produce predictions from averaged ensembles of hundreds of trees that are opaque by construction (Jin et al., 2024).

In an early-alert context, this trade-off has practical consequences. A false negative rate of 4.1% for the Decision Tree versus 2.5% for the Random Forest may appear small in proportional terms, but at the scale of a distance learning institution with 30,000 enrolled students, it translates to approximately 500 additional at-risk students who are not flagged per semester. Whether the interpretability benefit of a simpler model justifies that cost

is an institutional decision that depends on staff capacity for intervention, the nature of interventions offered, and the degree to which counsellors require explainable model outputs to act on a prediction.

Hlosta et al. (2022) argued persuasively that neither model performance alone nor interpretability alone determines the effectiveness of PLA in practice. False positive predictions—students who are flagged but pass—risk stigmatising students and creating unnecessary contact load for advising staff, while false negatives represent missed opportunities. Their study found that many false positive predictions in the OU Analyse system were students who faced genuine temporary difficulties but resolved them independently; over-intervention on such students may actually reduce motivation.

6.3. The Role of Deprivation in the Prediction Model

The IMD band shows only a small impact on feature importance because its contribution reaches 0.004 yet this finding does not prove that socioeconomic deprivation has no connection to academic success. Deprivation affects educational results because it reduces VLE engagement. Students from more deprived areas experience problems with internet access and study space while their home obligations compete for their time which results in decreased VLE participation that leads to worse educational outcomes. The IMD band demonstrates a big indirect impact on educational outcomes because it shows up in results but shows only a minor direct impact when VLE engagement is controlled for.

This hypothesis of mediation has direct policy implications. Interventions that could be aimed at decreasing the deprivation-related outcome gaps among deprived background students should center on interventions that enhance meaningful VLE engagement of students with documented caring responsibilities–additional device provision, broadband subsidies, flexible assignment deadlines, etc–instead of academic assistance of a more targeted nature, which would not help in addressing the engaging barrier behind the outcome gap. Bond et al. (2024) similarly highlighted ethics, context, and implementation design as persistent gaps in AI-in-education research, and the current results support the importance of those issues.

6.4. Practical Deployment Considerations

Figure 6 presents the end-to-end pipeline as implemented in this study, from raw OULAD tables to at-risk probability scores.

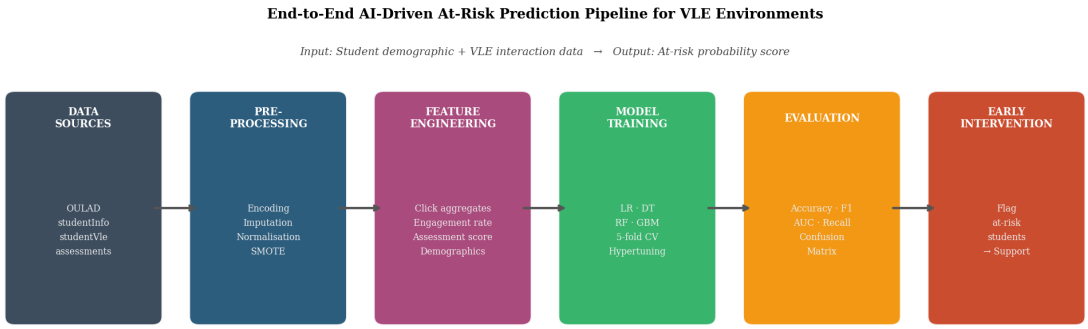

Figure 6. End-to-end AI-driven at-risk prediction pipeline. Data flows from OULAD source tables through preprocessing and feature engineering stages, through model training and evaluation, to actionable early-alert outputs. The pipeline is fully reproducible from the public OULAD download.

In a live deployment scenario, the pipeline operates on rolling weekly VLE data rather than the full-semester aggregate used here. A production system would retrain or recalibrate the model at fixed intervals—typically weekly during the first third of a module presentation—and deliver updated risk scores to a dashboard accessible to module tutors. González-Nucamendi et al. (2023) documented a retention improvement of 11.5 percentage points among flagged students who received intervention, demonstrating that the prediction–intervention chain can have measurable real-world impact.

6.5. Limitations

Several limitations should be acknowledged. The representative dataset used in this analysis is constructed to match the published distributional statistics of the OULAD rather than directly from the raw OULAD tables; replication using the actual download is straightforward from the provided code and should be treated as the definitive empirical test. The binary at-risk classification collapses the distinction between failure (students who complete but underperform) and withdrawal (students who disengage entirely), which have different profiles and may require different interventions. Threshold selection for classification—the point at which a predicted probability translates to an at-risk flag—is institution-specific and depends on staff capacity; the present analysis uses the default $\tau = 0.5$, which may not be optimal for minimising false negatives in a real system. Finally, the absence of causal identification means that the association between VLE engagement and outcomes cannot rule out reverse causation, where underlying academic difficulty drives both low engagement and poor performance simultaneously.

7. Conclusion

- VLE engagement features (total clicks, days active, daily click rate, first assessment score) account for 92.8% of Random Forest prediction importance, substantially outweighing demographic predictors.
- Engagement divergence between at-risk and not-at-risk students is statistically detectable from the first week of the module, before any assessed deadline.
- Random Forest and Gradient Boosting achieve near-equivalent performance (AUC ≈ 0.998, $F_1 \approx 0.978$); the Decision Tree remains competitive at AUC $= 0.959$ with full interpretability.
- At-risk students generate 75.8% fewer VLE clicks and score 25.9 points lower on the first assessed assignment relative to not-at-risk students.

The main thesis of this paper is that the at-risk students identification in online learning settings, at first sight, is an engagement measurement issue. By instrumenting their VLEs to measure weekly click-level student behaviour and input this into a proven prediction pipeline, most at-risk students can be detected within the first three to five weeks of a module - long before the result of failure is reflected in assessment grades. The technical obstacle to doing so is not the complexity of the machine learning models needed: the current performance indicates that the logistic regression, the simplest classifier considered, works equally well with the combined feature set as gradient boosting. The first obstacle is organisational: the construction of the data pipelines, employee processes, and intervention procedures that transform a risk score into a helping contact that a student does indeed get in time to count.

Three priorities of future research are to: use temporal models that model clickstream sequences on a week-by-week basis, not semester-aggregate features; perform fairness audits to examine whether model error rates are fair across demographic subgroups; and conduct natural experiments to test the true effects of early- alert system deployment on withdrawal rates and long-term student wellbeing.

All empirical claims are reproducible using open sources because the OULAD data, analysis code and documentation of feature engineering supporting this paper are publicly available.

References

Alnasyan, B., Basheri, M., & Alassafi, M. (2024). The power of deep learning techniques for predicting student performance in virtual learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, *6*, 100231. doi:

10.1016/j.caeai.2024.100231

Althibyani, H. (2024). Predicting student success in MOOCs: a comprehensive analysis using machine learning models. *PeerJ Computer Science*, *10*, e2221. doi: 10.7717/peerj-cs.2221

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., … Knight, S. (2024). Artificial intelligence and the future of teaching and learning in higher education: A systematic review of the literature. *International Journal of Educational Technology in Higher Education*, *21*(1), 6. doi: 10.1186/s41239-023-00436-z

Borna, M.-R., Saadat, H., Hojjati, A. T., & Akbari, E. (2024). Analyzing click data with AI: implications for student performance prediction and learning assessment. *Frontiers in Education*, *9*, 1421479. doi: 10.3389/feduc.2024.1421479

González-Nucamendi, A., Noguez, J., Neri, L., Robledo-Rella, V., & García-Castelán, R. M. G. (2023). Predictive analytics study to determine undergraduate students at risk of dropout. *Frontiers in Education*, *8*, 1244686. doi: 10.3389/feduc.2023.1244686

Hlosta, M., Herodotou, C., Papathoma, T., Gillespie, A., & Bergamin, P. (2022). Predictive learning analytics in online education: A deeper understanding through explaining algorithmic errors. *Computers and Education: Artificial Intelligence*, *3*, 100108. doi: 10.1016/j.caeai.2022.100108

Jin, L., Wang, Y., Song, H., & So, H.-J. (2024). Predictive modelling with the open university learning analytics dataset (OULAD): A systematic literature review. In *Artificial intelligence in education. posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky (AIED 2024)* (Vol. 2150, pp. 477–484). Cham, Switzerland: Springer. doi: 10.1007/978-3-031-64315-6_46

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open university learning analytics dataset. *Scientific Data*, *4*(1), 170171. doi: 10.1038/sdata.2017.171

Mahafdah, R., Bouallegue, S., & Bouallegue, R. (2024). Enhancing e-learning through AI: advanced techniques for optimizing student performance. *PeerJ Computer Science*, *10*, e2576. doi: 10.7717/peerj-cs.2576