ASCILITE 2026

The Future is Connection:

Collaboration, Creativity and Compassion in a Digital Society

Beyond counting clicks: Using the Resource Diversity Index (RDI) to improve early academic risk prediction in Virtual Learning Environments

Levi Arinze Ugwu, Kranthi Addanki, Euijoon Ahn
James Cook University

Student attrition remains a persistent challenge in higher education, with institutions reporting substantial dropout rates and many at-risk students failing to receive timely intervention. Existing learning analytics models predominantly rely on click frequency and interaction volume where models count interactions but do not adequately interpret the quality, timing, and diversity of engagement. This paper uses the Resource Diversity Index (RDI), a novel behavioural engagement feature that measures how broadly students distribute their interactions across different Virtual Learning Environment (VLE) activity types. Using the Open University Learning Analytics Dataset (OULAD), which contains records for 32,593 students across 22 courses, RDI was evaluated alongside Engagement, Regularity, and Anti-Procrastination features within an Extreme Gradient Boosting (XGBoost) classification framework using stratified five-fold cross-validation. Preliminary results show that adding RDI improved Accuracy from 0.8303 to 0.8357, AUC-ROC from 0.9014 to 0.9068, and F1-score from 0.7852 to 0.7918. These pilot findings indicate that the click-distribution diversity captured by RDI provides meaningful predictive information beyond interaction volume alone and offers a promising direction for more interpretable and proactive academic risk prediction.

Keywords: Academic risk prediction, learning analytics, resource diversity index, behavioural engagement, XGBoost, explainable AI

1. Introduction

Student retention and academic success remain major challenges for higher education institutions. Across OECD countries, only 43% of students who begin a bachelor's programme complete it within the theoretical duration, rising to 59% after one additional year and 70% after three, meaning roughly three in 10 entrants still have not completed even several years past the expected timeline (OECD, 2025). Conventional academic warning systems rely on end-of-semester grades, delaying opportunities for timely intervention (Chang et al., 2025). Meanwhile, Virtual Learning Environments (VLEs) such as Moodle, Blackboard, and Canvas generate rich clickstream data that could support earlier identification of at-risk students. However, a systematic review by Duro et al. (2026) found that most risk-prediction studies rely primarily on static academic records, with only a minority incorporating behavioural interaction data, highlighting the need to rebalance predictive algorithms toward behavioural signals.

A key limitation of existing models is their reliance on interaction counts rather than behavioural interpretation, a problem termed Counting Bias. For example, two students may each record 450 VLE clicks: one engages consistently throughout the semester, while the other concentrates activity in the 48 hours before an exam. Traditional click-volume models treat both identically despite their differing risk profiles. Although XGBoost identified sum_clicks as a dominant predictor in the OULAD dataset (Adnan et al., 2021), this metric captures neither timing, consistency, nor diversity of engagement.

This research-in-progress study addresses these gaps by designing and evaluating a behavioural feature engineering framework comprising Engagement (E), Regularity (R), Anti-Procrastination (AP) (Alhazbi, 2026), and a proposed Resource Diversity Index (RDI). Prior entropy and diversity-based measures have primarily characterised temporal regularity (Mai et al., 2023) or learning-tactic detection (Jovanović et al., 2017), rather than the breadth of engagement across VLE resource types that RDI is designed to capture. The study asks:

RQ1: How can behavioural features be engineered from VLE logs to capture risk beyond raw metrics?
RQ2: To what extent do these features improve predictive accuracy?

2. Theoretical framework

The proposed framework is grounded in four complementary learning theories. First, Self-Regulated Learning (SRL) (Zimmerman, 2002) explains how successful students cycle through forethought, performance control, and self-reflection, providing the rationale for Regularity (R) and Anti-Procrastination (AP), which capture consistent engagement and timely submission. Second, multidimensional Engagement theory (Fredricks et al., 2004) frames engagement across behavioural, cognitive, and affective dimensions, underpinning Engagement (E) and its three sub-indicators: content (E1), assessment (E2), and social engagement (E3). Third, Constructivist Learning Theory (Banihashem et al., 2022) supports E3 by positioning peer interaction and collaboration as essential to academic persistence. Fourth, Agentic Engagement Theory (Reeve, 2013) extends the tripartite model (Fredricks et al., 2004) by positioning students as proactive shapers of their learning environments, operationalised through the RDI which measures how broadly and evenly interactions are distributed across VLE activity types using Shannon entropy (H) (Shannon, 1948). Together, E, R, and AP constitute Alhazbi's (2026) composite engagement score, extended with RDI as a fourth dimension, capturing quality, consistency, timeliness, and breadth beyond raw click-count baselines.

3. Methods

3.1 Dataset: Open University Learning Analytics Dataset (OULAD)

This study utilized the OULAD, a publicly available and anonymized dataset containing records of 32,593 students across 22 courses (Kuzilek et al., 2017). OULAD includes seven non-relational core tables, but only four tables were selected and used for the analysis conducted in this study:

- studentInfo which contains student demographic data.
- studentAssessment contains data describing their individual assessment scores and submission dates.
- studentVle includes time-stamped clickstream logs of all interaction with OU VLE activities, including total number of clicks.
- vle comprises of metadata assigning an OU VLE activity identifier to the type of activity (13 different types) and to the week or module associated with that identifier.

The binary target classifies students as At Risk (1) or Not At Risk (0), derived from OULAD's final_result field: Withdrawn/Fail = At Risk; Pass/Distinction = Not At Risk. Class imbalance is addressed via stratified k-fold cross-validation and XGBoost class weighting, with SMOTE applied only to each fold's training portion to prevent data leakage. Table 1 classifies OULAD's 13 VLE activity types into three functional categories used in feature engineering. Predictions use cumulative VLE and assessment data up to the end of week four; later clicks, scores, and submissions are excluded to prevent temporal leakage.

Table 1.
Classification of OULAD VLE activity types into functional categories

Category	OULAD activity types	Description
Content	Page, Resource, URL, Oucontent, Subpage	Passive course materials students read or download
Assessment	Quiz, Ouelluminate, Questionnaire, Externalquiz	Evaluative activities students complete or submit
Social	Forumng, Wiki, Glossary, Collaborate	Peer interaction and shared knowledge-building tools

3.2 Behavioural engagement feature engineering model

This is a structured sequence of processing steps that transforms raw behavioral engagement data into structured features for machine learning analysis and prediction. The four theory-grounded indicators as described in Section 2 are

derived from OULAD clickstream data. Table 2 summarises each indicator's formula, theoretical grounding, and value interpretation.

Table 2.
Behavioural engagement indicator

Indicator	Formula	Description	Reference
Engagement (E)	$Ex = \Sigma C_{s,c,t} / \Sigma T_{c,t}$ x ∈ {total, content, assessment, social}	$C_{s,c,t}$: Total clicks by student s in course c for activity type t $T_{c,t}$: Total course-level c clicks for activity type t. Range: 0 (no engagement) → 1 (proportional to course average)	Fredricks et al. (2004); Banihashem et al. (2018)
Regularity (R)	$R_{entropy} = H/\log(n)$; $H = -\Sigma (p_i \log p_i)$ $R_{active} = N_{active}/N_{total}$	N_{active} : Weeks with at least one recorded interaction. N_{total} : Total number of weeks. H : Shannon entropy of weekly activity distribution. pi = wi / Σwj (for j = 1..n weeks), Range: 0 (fully irregular/cramming) → 1 (perfectly consistent)	Zimmerman (2002)
Anti-Procrastination (AP)	$AP(s,c) = (1/N_a)\times\Sigma(T_{s,a})$ $T_{s,a} = (D_a - S_{s,a})/(D_a - S_{min,a})$	D_a : Assessment due date (days from course start) $S_{s,a}$: Student submission $S_{min,a}$: Earliest cohort submission $T_{s,a}$: Assessment Submission Timeline Range: AP ≈ 0 (last-minute/risk) → AP ≈ 1 (early/proactive)	Zimmerman (2002)
Resource Diversity Index (RDI)	RDI = H/log(k); K = 13 fixed) $H = -\Sigma (p_i \log p_i)$	p_i: clicks on activity type i / total clicks; k: number of VLE activity types available (fixed at 13 activity types, per Table 1). H: Shannon entropy of interaction distribution across the activity types. 0 (all clicks on a single type) → 1 (uniform across all k = 13 activity types)	Reeve (2013); Shannon (1948);

3.3 Composite behavioural engagement score model

The Composite Behavioural Engagement Score (B_i) is a multidimensional metric representing students' overall learning behaviour and academic risk, using behavioural indicators extracted from OULAD. Building on the three-dimensional baseline from section two (Alhazbi, 2026), where E carries 64% weight, R 20%, and AP 16%:

Baseline: $B_i = 0.64 \cdot E + 0.20 \cdot R + 0.16 \cdot AP$
Proposed: $B_i = w_1 \cdot E + w_2 \cdot R + w_3 \cdot AP + w_4 \cdot RDI$

This study extends the baseline by incorporating RDI as a fourth dimension. Weights w_1–w_4 are not fixed a priori but derived from XGBoost feature importance and SHAP mean absolute values through the M4 ablation run (Table 3), reflecting each dimension's predictive contribution. $B_i \in [0, 1]$, where 1 indicates lower risk and 0 indicates higher risk.

3.4 Predictive modelling

Extreme Gradient Boosting (XGBoost) combines multiple decision trees sequentially, correcting errors at each step to produce accurate classifications (Chen & Guestrin, 2016). It was selected for four reasons: superior performance on structured tabular data; native handling of missing VLE values without imputation; mathematical compatibility with TreeSHAP for exact Shapley value computation; and demonstrated dominance over Support Vector Machine and Random Forest on OULAD (Alhazbi, 2026). Stratified five-fold cross-validation with Optuna Bayesian hyperparameter

optimisation ensures valid performance estimates, while SHapely Additive exPlanation (SHAP) provides individual-level feature attributions.

3.5 Evaluation

The framework is evaluated using an ablation study. New behavioural features are added step by step which allows for the individual measurement of each feature as summarized in Table 3. Performance is assessed on Accuracy, Precision, Recall, F1-Score (macro/weighted), and AUC-ROC. Paired t-tests across folds determine statistical significance.

Table 3.
Ablation study specifications

Model	Label	Feature set	Purpose
M1	Baseline	Raw click counts + demographic variables + raw assessment scores	Current state-of-practice; establishes Counting Bias benchmark
M2	Behavioural	M1 + E_total, E_content, E_assess, E_social, R_entropy, R_active, AP	Replicates Alhazbi (2026) feature set
M3	Diversity	M2 + Resource Diversity Index (RDI = H/log(k),k = 13)	Isolates RDI marginal contribution
M4	Composite	M3 + composite Bi = $w_1 \cdot E + w_2 \cdot R + w_3 \cdot AP + w_4 \cdot RDI$	Evaluates composite score as a derived meta-feature

4. Preliminary results

Two XGBoost models were trained and evaluated on OULAD using stratified five-fold cross-validation. Feature engineering for all four indicators has been fully implemented in Python (Pandas, NumPy, Scikit-learn). Table 4 reports results as mean ± standard deviation across folds.

Table 4.
Preliminary model performance: XGBoost on OULAD (5-fold CV)

Model	Accuracy	AUC-ROC	F1-Score	Features
XGBoost Baseline (Alhazbi, 2026)	0.8303 ± 0.0034	0.9014 ± 0.0034	0.7852 ± 0.0079	E+R+AP
XGBoost + RDI (Proposed)	**0.8357 ± 0.0065**	**0.9068 ± 0.0053**	**0.7918 ± 0.0064**	**E+R+AP+RDI**

The bolded values indicate the RDI-augmented model. These results confirm the establishment of Alhazbi's (2026) benchmark and demonstrate that adding RDI yields consistent improvements across all three metrics: Accuracy +0.54 percentage points (pp), AUC-ROC +0.54pp, F1-Score +0.66pp. These gains suggest, but not yet statistically confirm, that click-distribution diversity may provide predictive information beyond existing behavioural indicators, pending completion of the planned paired t-tests and M3 and M4 ablation analysis.

5. Discussion and implications

Preliminary results indicate that behavioural diversity is a significant dimension of student engagement. Unlike traditional analytics, which measure how much students interact with the VLE, RDI captures how students distribute interaction across learning resources. Grounded in Agentic Engagement Theory, RDI is derived entirely from existing VLE logs, enabling integration into current learning analytics systems without additional data collection or proprietary technologies.

In future, a GenAI explanation layer could address the interpretability gap by translating SHAP feature attributions into natural-language explanations, helping educators understand why a student is at risk and supporting targeted interventions. Key limitations include generalisability, as findings are currently limited to the OULAD dataset, and algorithmic fairness, since RDI may disadvantage students with limited VLE access due to disability, lower Index of Multiple Deprivation, or connectivity issues. RDI-based risk flags should therefore be interpreted alongside contextual access information, not as a sole determinant of intervention.

6. Conclusion

This research-in-progress study uses the RDI as a novel fourth dimension of the behavioural engagement framework, extending three indicator model (E, R, AP) to address the Counting Bias prevalent in academic risk prediction. By operationalising Shannon entropy across OULAD's 13 VLE activity types, RDI captures the breadth of a student's learning strategy: a dimension complementary to and not captured by existing behavioural indicators. Preliminary XGBoost results demonstrate that incorporating RDI into the composite engagement score (B_i) produces consistent improvements in Accuracy, AUC-ROC, and F1-Score, confirming that click-distribution diversity provides meaningful incremental predictive value for student academic risk prediction. Future work can integrate SHAP-based explainability and evaluate GenAI-generated risk narratives with academic advisors.

References

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Bashir, 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. https://doi.org/10.1109/ACCESS.2021.3049446

Alhazbi, S. (2026). Pedagogically-informed behavioural learning analytics: An expert approach to predicting at-risk students. *Expert Systems, 43*(3), e70210. https://doi.org/10.1111/exsy.70210

Banihashem, S. K., Farrokhnia, M., Badali, M., & Noroozi, O. (2022). The impacts of constructivist learning design and learning analytics on students' engagement and self-regulation. *Innovations in Education and Teaching International*, *59*(4), 442–452. https://doi.org/10.1080/14703297.2021.1890634

Chang, Y. H., Chen, F. C., & Lee, C. I. (2025). Developing an early warning system with personalized interventions to enhance academic outcomes for at-risk students in Taiwanese higher education. *Education Sciences, 15*(10), 1321. https://doi.org/10.3390/educsci15101321

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Duro, B., Gomes, A., Correia, F. B., Borges, A. R., & Bernardino, J. (2026). Machine learning and deep learning for dropout prediction in higher education: A review. *Computers, 15*(3), 164. https://doi.org/10.3390/computers15030164

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Hassan, S. U., Waheed, H., Aljohani, N. R., Ali, M., Ventura, S., & Herrera, F. (2019). Virtual learning environment to predict withdrawal by leveraging deep learning. *International Journal of Intelligent Systems, 34*(8), 1935–1952. https://doi.org/10.1002/int.22129

Jovanović, J., Gašević, D., Dawson, S., Pardo, A., & Mirriahi, N. (2017). Learning analytics to unveil learning strategies in a flipped classroom. *The Internet and Higher Education, 33*, 74–85. https://doi.org/10.1016/j.iheduc.2017.02.001

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

Mai, T. T., Crane, M., & Bezbradica, M. (2023). Students' learning behaviour in programming education analysis: Insights from entropy and community detection. *Entropy, 25*(8), 1225. https://doi.org/10.3390/e25081225

OECD. (2019*). Education at a glance 2019: OECD indicators*. OECD Publishing. https://doi.org/10.1787/f8d7880d-en

Reeve, J. (2013). How students create motivationally supportive learning environments for themselves: The concept of agentic engagement. *Journal of Educational Psychology, 105*(3), 579–595. https://doi.org/10.1037/a0032690

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice, 41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Ugwu, L., Addanki, K., & Ahn, E. (2026). Beyond counting clicks: Using the Resource Diversity Index (RDI) to improve early academic risk prediction in Virtual Learning Environments. In Cochrane, T., Narayan, V., Poole, T., Lai, P., Lake, P., Moore, O., & Li, L. (Eds.), The Future is Connection: Collaboration, Creativity and Compassion in a Digital Society. Proceedings ASCILITE 2026. Brisbane (pp. 860–865). https://doi.org/10.65106/apubs.2026.3271

Note: All published papers are refereed, having undergone a double-blind peer-review process.
The author(s) assign a Creative Commons by attribution license enabling others to distribute, remix, tweak, and build upon their work, even commercially, as long as credit is given to the author(s) for the original creation.

© Ugwu, L., Addanki, K., & Ahn, E. 2026