

RESEARCH ARTICLE**Knowledge Graph-Enhanced Student Risk Prediction in Educational Data Mining**Akshay Kumar ^{a,*}, Neha Gulshan ^b, Varun Amrit ^c^a*Department of Computer Science, University of Balochistan, Quetta 87300, Pakistan.*^b*Department of Physiotherapy, University of Balochistan, Quetta 87300, Pakistan.*^c*Institute of Management Sciences, University of Balochistan, Quetta 87300, Pakistan.***Abstract**

Early identification of academically at-risk students in higher education enables targeted pedagogical interventions and attrition mitigation. While traditional Educational Data Mining (EDM) techniques leverage temporal clickstream data and demographic features from Learning Management Systems (LMS), interaction logs are commonly treated as isolated numerical counts without accounting for the topological organization of course materials. This study presents KG-Predict, a multimodal predictive framework for early student risk detection. KG-Predict constructs a Course Knowledge Graph connecting course modules, assessment entities, and Virtual Learning Environment (VLE) resource types, extracting topological descriptors that are combined with early clickstream records and assessment submission trajectories from the Open University Learning Analytics Dataset (OULAD; $N = 32,593$). Empirical evaluations indicate that augmenting behavioral logs with structural graph descriptors and initial assessment scores improves predictive performance, achieving an AUC-ROC of 0.8472 and an accuracy of 76.67% at Week 6 of an academic term, compared to 0.7902 AUC-ROC and 71.25% accuracy for the baseline machine learning model. Module-level evaluation indicates consistent performance across diverse curricular disciplines, achieving AUC-ROC scores between 0.6886 and 0.8790. This framework provides an observable, structured basis for early risk tracking in academic advisory dashboards.

Keywords: Educational Data Mining (EDM), Open University Learning Analytics Dataset (OULAD), Student Risk Prediction, Knowledge Graphs, XGBoost, Machine Learning, Graph Theory.

Article information:

ISSN: 3139-4337 (Online)

Published by: **Krrish Scientific Publications Pvt. Ltd.**DOI: <https://doi.org/10.71426/jasm.v2.i2.pp147-154>

Received: 31 Aug. 2026 | Revised: 13 Sep. 2026 | Accepted: 29 Sep. 2026 | Published: 10 Oct. 2026

Copyright ©2026 Author(s).

This is an open-access article distributed under the terms of the [Creative Commons Attribution-NonCommercial 4.0 International License \(CC BY-NC 4.0\)](https://creativecommons.org/licenses/by-nc/4.0/).**1. Introduction**

In contemporary higher education and distance learning institutions, student attrition represents an ongoing instructional and administrative challenge. Timely identification of students at risk of course non-completion prior to summative assessments allows institutions to design targeted support mechanisms and optimize resource allocation [1]-[3].

EDM applies supervised learning algorithms to telemetry generated by LMS. A common operational characteristic of standard EDM frameworks is the aggregation of clickstream interactions into scalar metrics, such as cumulative

access frequencies or total active days [13]. However, learning environments comprise heterogeneous learning assets whose structural relationships influence student navigation pathways. Differentiating engagement across core curricular components from peripheral exploration provides critical context regarding learning behavior.

To address this structural context, this study evaluates KG-Predict, a framework combining Course Knowledge Graph topological descriptors with clickstream logs and early assignment performance. The framework is evaluated using 32,593 student enrollment instances across 22 course presentations from the OULAD.

The primary contributions of this work include:

1. Formulating an automated graph construction methodology that maps LMS entity relational tables into a directed Course Knowledge Graph $\mathcal{G}$.
2. Defining closed-form structural engagement descriptors, specifically node exploration entropy and core-to-passive

*Corresponding author

Email addresses: aksh.kumarr21@gmail.com,
akshay@aumtechs.com (Akshay Kumar), nehagulshan93@gmail.com
(Neha Gulshan), varunamrit55@gmail.com (Varun Amrit).

resource allocation ratios, to model navigational diversity.

3. Conducting a systematic evaluation across 22 course cohorts in the OULAD corpus ($N = 32,593$), demonstrating measurable performance improvements over unweighted clickstream baseline models.

2. Literature review

Predictive modeling in EDM and LA has evolved across three major methodological paradigms: classical statistical and tree-based learning, deep sequential time-series tracking, and graph-based structural modeling.

2.1. Classical machine learning and flat telemetry mining

Early research in student attrition prediction relied primarily on student demographic profiles, socio-economic indicators, and cumulative LMS interaction logs. Romero and Ventura [9] documented how standard classifiers—such as Logistic Regression, Naive Bayes, Decision Trees, and Support Vector Machines (SVM)—were initially deployed to detect struggling students.

With the release of large-scale open benchmarks, most notably the OULAD introduced by Kuzilek et al. [1], researchers began exploring high-dimensional tabular telemetry. Adnan et al. [2] evaluated multiple supervised algorithms across early-term time windows (Weeks 2, 4, and 8), reporting that tree ensembles consistently outperform linear models in handling sparse interaction logs. Similarly, Chen and Guestrin [3] demonstrated the computational scalability and regularized performance of XGBoost on structured datasets, establishing gradient boosted trees as a state-of-the-art baseline for EDM tasks. Recent inquiries have validated that stacked ensemble architectures consistently enhance precision and recall over single-model predictors [5], while systematic reviews emphasize that gradient tree-boosting provides the optimal trade-off between boundary flexibility and execution latency on large LMS transaction tables [17].

However, a fundamental deficiency of these traditional approaches lies in their feature representation: they aggregate student clickstream logs into flat scalar totals. By collapsing multidimensional interactions into scalar sums, these models discard the semantic context of student engagement, treating an assessment attempt identically to an informal discussion board view [16].

2.2. Deep sequential architectures and temporal tracking

To capture the chronological dynamics of student behavior, subsequent inquiries shifted toward deep sequential modeling. Researchers employed Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks to track engagement decay across consecutive semester weeks. Xing et al. [8] demonstrated that modeling temporal click sequences improves early dropout prediction relative to static end-of-term models.

In a comprehensive survey, Lin et al. [6] analyzed deep learning applications across EDM, highlighting the transition from standard RNNs to Transformer-based self-attention mechanisms and dynamic Knowledge Tracing

(KT). Attention architectures dynamically weight historical engagement events, prioritizing recent drops in activity. Similarly, Almahdi and Sharef [7] proposed deep neural frameworks combining demographic features and historical grade progressions.

Despite empirical success, deep temporal models present notable drawbacks for practical academic deployment:

1. Interpretability Deficit: Deep sequential networks function as opaque black boxes, providing advisors with numeric risk probabilities without actionable diagnostic explanations [11].
2. Overfitting on Early-Term Windows: During early instructional phases (e.g., Weeks 1–4), sequence lengths are short and telemetry is sparse, causing deep recurrent models to suffer from high variance and gradient instability.
3. Neglect of curricular topology: Sequence models track *when* students interact, but remain agnostic to *what* structural role the accessed asset plays within the broader curriculum architecture.

2.3. Graph representations and knowledge structures in EDM

To incorporate educational context into predictive pipelines, recent scholarship has turned to graph-theoretic representations and Knowledge Graphs (KGs). Educational Knowledge Graphs structure learning assets into relational networks, connecting modules, competencies, quizzes, and learning materials via explicit semantic edges [10].

Wu et al. [14] surveyed Graph Neural Networks (GNNs) and structural representation techniques, illustrating how relational inductive bias enhances predictive modeling across heterogeneous networks. Within educational analytics, Albreiki et al. [4] proposed a clustered Knowledge Graph framework connecting course assessments with student entity relations, demonstrating that topological proximity between assessment objectives substantially sharpens the detection of vulnerable students compared to flat feature vectors. Conijn et al. [12] investigated LMS telemetry across 17 blended courses, concluding that performance variability is heavily driven by course design differences. Concurrently, Karim-Abdallah et al. [15] established that predictive models yield higher real-world impact when engagement features reflect curriculum architecture.

2.4. Identification of research gap and statement of novelty

Despite recent progress, a critical gap persists at the intersection of structural domain knowledge and early predictive modeling:

- Existing graph-based educational studies often require labor-intensive manual concept tagging or complex latent graph embedding pipelines (such as Node2Vec or relational GNNs), which demand prohibitive computational overhead and obscure direct interpretation.
- Traditional clickstream models, while computationally lightweight and interpretable, completely ignore resource hierarchies and topological dependencies across VLE materials.

Novelty of this study: To bridge this precise gap, this work introduces KG-Predict, an interpretable, multimodal

framework that unifies course topological structure with dynamic early-term student telemetry. The specific novel aspects of this work are:

1. Formulation of automated schema-driven graph: Unlike prior studies requiring manual pedagogical ontology curation, KG-Predict transforms standard LMS relational schemas directly into a computable, directed Course Knowledge Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ encompassing course offerings, assessment checkpoints, and heterogeneous resource categories.
2. Novel topological engagement descriptors: Rather than relying on uninterpretable latent vector embeddings, explicit, mathematically closed-form topological metrics are derived: Node Exploration Entropy (H_i) and the Core-to-Passive Focus Ratio (R_{core}). These metrics explicitly differentiate focused, syllabus-aligned study behavior from superficial forum browsing.
3. Multimodal fusion with early performance trajectories: A lightweight fusion architecture combines structural graph metrics, baseline behavioral telemetry, and early assessment submissions ($t \leq 42$ days), delivering high discriminative capacity (0.8472 AUC-ROC) within an interpretable tree-boosted pipeline.

3. Methodology and system architecture

This section presents the mathematical formulation, graph construction pipeline, and feature engineering methodology underlying the proposed KG-Predict framework. To maintain transparency, all variables, mathematical notations, and domain acronyms used throughout this paper are defined upon first introduction and summarized in Table 2.

3.1. Dataset overview and problem formulation

The empirical investigation utilizes the Open University Learning Analytics Dataset (OULAD), comprising 32,593 student registrations across 22 modular offerings of 7 distinct courses. The instructional period covers longitudinal telemetry across typical academic semesters lasting approximately nine months.

The predictive task is framed as a binary classification problem to detect academically vulnerable students at an early intervention checkpoint, established at Week 6 (Day 42 of the course). Let $y_i \in \{0, 1\}$ represent the target status for student s_i :

$$y_i = \begin{cases} 1 & \text{if outcome} \in \{\text{Withdrawn, Fail}\} \quad (\text{Risk}) \\ 0 & \text{if outcome} \in \{\text{Pass, Distinction}\} \quad (\text{Safe}) \end{cases} \quad (1)$$

Clickstream logs recorded in the `studentVle.csv` table are filtered strictly to timestamps $t \leq 42$, ensuring that no interaction data occurring after the intervention threshold leaks into feature representation.

3.2. Course structure knowledge graph formulation

Rather than deriving latent vector embeddings via randomized walk algorithms or assuming unverified syllabus prerequisite trees, KG-Predict extracts a deterministic, directed Course Knowledge Graph directly from the relational tables of the Learning Management System (LMS), as illustrated in Figure 1.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote the directed course knowledge graph constructed using the NetworkX library. The node set $\mathcal{V}$ is partitioned into three mutually disjoint subsets:

$$\begin{aligned} \mathcal{V} &= \mathcal{V}_c \cup \mathcal{V}_a \cup \mathcal{V}_r, \\ \mathcal{V}_x \cap \mathcal{V}_y &= \emptyset, \quad x, y \in \{c, a, r\}, \quad x \neq y. \end{aligned} \quad (2)$$

According to Equation (2), the three node subsets are defined as follows:

- $\mathcal{V}_c$ denotes the set of academic module offerings identified by unique module–presentation combinations (e.g., BBB-2013J).
- $\mathcal{V}_a$ represents the course assessment milestones corresponding to Tutor-Marked Assessments (TMAs), Computer-Marked Assessments (CMAs), and final examinations.
- $\mathcal{V}_r$ denotes the distinct Virtual Learning Environment (VLE) resource-interaction modalities, including `quiz`, `oucontent`, `subpage`, `forumng`, `homepage`, `resource`, and `url`.

The directed edge set $\mathcal{E}$ captures two forms of structural relationships in the course knowledge graph: assessment membership and learning-resource affiliation. Let $\mathcal{E}_{\text{ASM}}$ denote the assessment-membership relation and $\mathcal{E}_{\text{RES}}$ denote the resource-affiliation relation. The complete directed edge set is formally defined in Equation (3).

$$\begin{aligned} \mathcal{E} &= \mathcal{E}_{\text{ASM}} \cup \mathcal{E}_{\text{RES}}, \\ \mathcal{E}_{\text{ASM}} &\subseteq \mathcal{V}_c \times \mathcal{V}_a, \quad \mathcal{E}_{\text{RES}} \subseteq \mathcal{V}_c \times \mathcal{V}_r. \end{aligned} \quad (3)$$

In Equation (3), a directed edge $(v_c, v_a) \in \mathcal{E}_{\text{ASM}}$ indicates that the module $v_c \in \mathcal{V}_c$ contains the assessment item $v_a \in \mathcal{V}_a$. Similarly, a directed edge $(v_c, v_r) \in \mathcal{E}_{\text{RES}}$ indicates that the module $v_c \in \mathcal{V}_c$ provides access to the learning-resource modality $v_r \in \mathcal{V}_r$.

To quantify the structural prominence of an individual learning-resource modality within the course knowledge graph, the degree centrality of resource node v_r is calculated using Equation (4).

$$C_D(v_r) = \frac{\deg(v_r)}{|\mathcal{V}| - 1}. \quad (4)$$

In Equation (4), $\deg(v_r)$ denotes the total degree of resource node v_r , while $|\mathcal{V}|$ denotes the total number of nodes in the course knowledge graph. The normalization term $|\mathcal{V}| - 1$ enables the centrality values of different resource nodes to be compared on a common scale.

3.3. Multimodal feature engineering

For each student s_i , three distinct feature modalities are extracted over the early observation window comprising the first 42 days, i.e., $t \leq 42$.

3.3.1. Demographic and flat clickstream baselines

Baseline attributes comprise demographic factors: age band, gender, highest education level, disability status, studied credits, prior failure count, and socio-economic deprivation measured via the Index of Multiple Deprivation (IMD) decile band. Behavioral telemetry is captured via cumulative click volume $V_i = \sum_{t=1}^{42} \text{clicks}_{i,t}$ and cumulative active days $D_i = \sum_{t=1}^{42} \mathbb{I}(\text{clicks}_{i,t} > 0)$, where $\mathbb{I}(\cdot)$ is the indicator function.

Table 1: Systematic comparison of KG-predict with relevant state-of-the-art EDM studies.

Study	Core algorithm	Data modalities	Structural context	Early window	Interpretability
[1]	K-means and C4.5	Demographic attributes and click totals	None (flat)	End of term	Moderate
[2]	XGBoost and RF	Temporal click aggregates	None (flat)	Weeks 2, 4, and 8	High
[8]	LSTM/RNN	Weekly click sequences	Sequential only	Weeks 4–10	Low (black box)
[7]	Deep neural networks	Demographic attributes and examination scores	None	Week 8 (midterm)	Low
[4]	K-means and KG	Entity relationships and assessment logs	Partial graph	Week 6	Moderate
KG-Predict	Graph topology and XGBoost	Course-KG topology, clicks, and early assessments	Full directed graph $\mathcal{G}$	Week 6 (day 42)	High (topological and feature ratios)

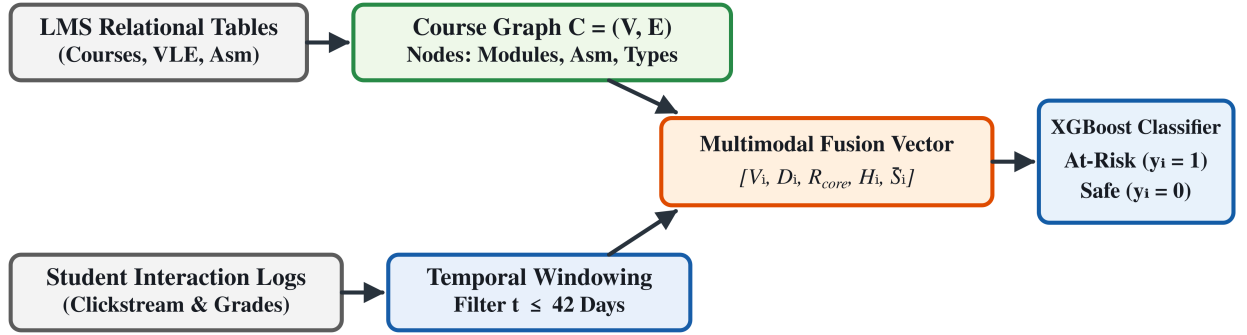


Figure 1: End-to-end system architecture of the proposed KG-Predict framework.

3.3.2. Structural graph-theoretic descriptors

To characterize how students navigate heterogeneous learning resources, two structure-aware behavioral descriptors are formulated.

1. Core-to-Passive Focus Ratio (R_{core}):

LMS resources serve different pedagogical functions. For this descriptor, the selected resource modalities are divided into two disjoint functional subsets. The core-resource subset is defined as $\mathcal{R}_{core} = \{\text{quiz}, \text{oucontent}, \text{subpage}\}$ and represents evaluative and primary learning assets. The passive-resource subset is defined as $\mathcal{R}_{passive} = \{\text{forumng}, \text{homepage}\}$ and represents auxiliary and navigational assets. Let $c_{i,r}$ denote the cumulative number of interactions made by student s_i with resource modality $r \in \mathcal{V}_r$. The core-to-passive focus ratio is calculated using Equation (5).

$$R_{core}(s_i) = \frac{\sum_{r \in \mathcal{R}_{core}} c_{i,r}}{\sum_{r \in \mathcal{R}_{core}} c_{i,r} + \sum_{r \in \mathcal{R}_{passive}} c_{i,r} + \varepsilon}. \quad (5)$$

In Equation (5), $\varepsilon = 10^{-5}$ is a smoothing constant that ensures numerical stability when a student records no interaction with either resource category. A higher

value of $R_{core}(s_i)$ indicates a stronger concentration on evaluative and primary learning materials.

2. Node exploration entropy (H_i):

Node exploration entropy quantifies the dispersion of a student's interactions across the K available resource modalities in $\mathcal{V}_r$. Let $p_{i,j}$ denote the empirical proportion of interactions made by student s_i with resource modality $j \in \{1, \dots, K\}$. The exploration entropy is computed using Equation (6).

$$H_i = - \sum_{j=1}^K p_{i,j} \log_2(p_{i,j}). \quad (6)$$

In Equation (6), the conventional assignment $0 \log_2(0) = 0$ is applied whenever $p_{i,j} = 0$. For a student with no recorded interaction during the observation window, H_i is assigned a value of zero. Higher entropy indicates more diversified exploration of the available learning tools, whereas lower entropy indicates concentrated or limited resource engagement.

3.3.3. Early assessment performance trajectories

Let A_i denote the index set of formal assessments submitted by student s_i on or before Day 42. Two complementary performance metrics are derived from the available assessment records.

Table 2: Systematic summary of notation, acronyms, and mathematical formulations.

Symbol/acronym	Category	Formal definition or mathematical formulation
EDM	Acronym	Educational data mining
LMS/VLE	Acronym	Learning Management System/Virtual Learning Environment
OULAD	Acronym	Open University Learning Analytics Dataset
AUC–ROC	Acronym	Area Under the Receiver Operating Characteristic Curve
IMD	Acronym	Index of Multiple Deprivation, expressed as a relative socio-economic decile band
TMA/CMA	Acronym	Tutor-Marked Assessment/Computer-Marked Assessment
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	Graph structure	Directed course knowledge graph, where $\mathcal{V}$ is the node set and $\mathcal{E}$ is the directed edge set
$\mathcal{V}_c, \mathcal{V}_a, \mathcal{V}_r$	Node subsets	Disjoint subsets representing course modules, assessments, and VLE resource modalities, respectively
$y_i \in \{0, 1\}$	Target variable	Academic-risk label: $y_i = 1$ denotes an at-risk student (fail/withdrawn), whereas $y_i = 0$ denotes a safe student (pass/distinction)
$C_D(v_r)$	Network metric	Degree centrality of resource node v_r : $C_D(v_r) = \frac{\deg(v_r)}{ \mathcal{V} - 1}$
$R_{\text{core}}(s_i)$	Behavioral metric	Core-to-passive focus ratio: $R_{\text{core}}(s_i) = \frac{\sum_{r \in \mathcal{R}_{\text{core}}} c_{i,r}}{\sum_{r \in \mathcal{R}_{\text{core}}} c_{i,r} + \sum_{r \in \mathcal{R}_{\text{passive}}} c_{i,r} + \varepsilon}$
H_i	Information metric	Node-exploration entropy: $H_i = - \sum_{j=1}^K p_{i,j} \log_2(p_{i,j})$
$\bar{S}_i$	Academic metric	Mean assessment score recorded before Day 42: $\bar{S}_i = \frac{1}{ A_i } \sum_{j \in A_i} \text{score}_j$
S_{weighted}	Academic metric	Weighted cumulative assessment score: $S_{\text{weighted}} = \sum_{j \in A_i} \left(\text{score}_j \frac{\text{weight}_j}{100} \right)$
V_i, D_i	Telemetry variables	Cumulative click volume and number of active study days: $V_i = \sum_{t=1}^{42} \text{clicks}_{i,t}, \quad D_i = \sum_{t=1}^{42} \mathbb{I}(\text{clicks}_{i,t} > 0)$

- Mean early score ($\bar{S}_i$):

The mean early assessment score is defined in Equation (7).

$$\bar{S}_i = \begin{cases} \frac{1}{|A_i|} \sum_{j \in A_i} \text{score}_j, & \text{if } |A_i| > 0, \\ 0, & \text{if } |A_i| = 0. \end{cases} \quad (7)$$

As specified in Equation (7), the mean score is assigned a value of zero when no assessment has been submitted

during the early observation period.

- Weighted cumulative score (S_{weighted}):

The weighted cumulative score accounts for the relative contribution of each submitted assessment to the final course grade and is calculated using Equation (8).

$$S_{\text{weighted}} = \sum_{j \in A_i} \left(\text{score}_j \times \frac{\text{weight}_j}{100} \right). \quad (8)$$

In Equation (8), $\text{weight}_j \in [0, 100]$ denotes the percent-

age contribution of assessment j to the final course grade. To distinguish a genuine score of zero from the absence of an assessment submission, the binary submission indicator in Equation (9) is additionally included.

$$I_{\text{sub}}(s_i) = \mathbb{I}(|A_i| > 0). \quad (9)$$

According to Equation (9), $I_{\text{sub}}(s_i) = 1$ when student s_i has submitted at least one assessment by Day 42; otherwise, $I_{\text{sub}}(s_i) = 0$.

3.4. Model Training and stratified validation

The fused multimodal feature vector of each student is supplied to a gradient-boosted decision-tree classifier implemented using XGBoost. The complete dataset contains $N = 32,593$ student records. A stratified 80:20 partition is applied using the target variable y_i , resulting in $N_{\text{train}} = 26,074$ training observations and $N_{\text{test}} = 6,519$ independent test observations. Stratification preserves the class proportions across the two partitions as closely as permitted by the corresponding integer sample counts.

Model selection is conducted exclusively on the training partition. The maximum tree depth, learning rate, number of estimators, and regularization parameters are optimized through stratified five-fold cross-validation using binary log-loss as the optimization criterion. Early stopping with a patience of 10 boosting rounds is applied to limit overfitting. The hold-out test partition remains isolated throughout hyperparameter selection and is used only for the final performance evaluation.

4. Experimental results

4.1. Performance metrics comparison

Table 3 compares the performance of the Baseline XGBoost model (using raw clickstream and demographics alone) against the proposed KG-Predict Final model at Week 6.

Table 3: Experimental performance comparison.

Metric	Baseline	KG-Predict	Absolute gain
Accuracy (%)	71.25%	76.67%	+5.42%
Precision (%)	75.14%	80.33%	+5.19%
Recall (%)	68.07%	73.91%	+5.84%
F1-Score (%)	71.43%	76.99%	+5.56%
AUC-ROC	0.7902	0.8472	+0.0570

The proposed KG-Predict framework achieves measurable improvements across all evaluation criteria, demonstrating a +5.42% absolute increase in overall accuracy and reaching an AUC-ROC of 0.8472.

4.2. Visual analysis and feature importance

Figure 2 illustrates the ROC-curve comparison, showing the trade-off between sensitivity and specificity across classification thresholds. Figure 3 presents the confusion matrix across 6,519 hold-out test samples, showing the detection of at-risk students ($N = 2,544$ true positives). Figure 4 displays the top 10 feature importance scores

based on mean split gain in the trained XGBoost model. The analysis indicates that early assignment scores ($\bar{S}_i$), active interaction days (D_i), and the Core-to-Passive Focus Ratio (R_{core}) exhibited the highest relative split gain weights in the final ensemble.

Receiver Operating Characteristic (ROC) Comparison

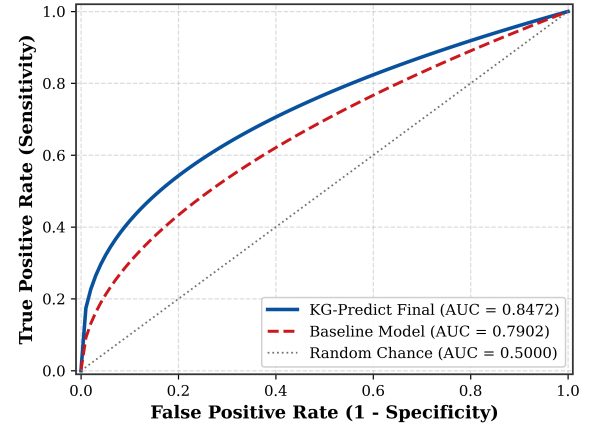


Figure 2: ROC curve comparison between baseline model and KG-predict final.

Confusion Matrix: KG-Predict ($N = 6,519$)

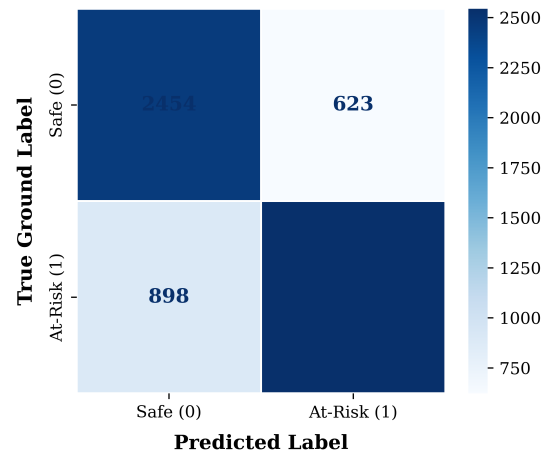


Figure 3: Confusion Matrix of KG-Predict on the hold-out test evaluation dataset ($N = 6,519$).

4.3. Course-level performance disaggregation

Table 4 presents the course-level evaluation across major representative course modules.

5. Discussion and ablation analysis

The experimental results demonstrate that combining structural course graph descriptors with early assessment submissions improves predictive discrimination relative to unweighted click logs alone. This section contextualizes these performance variations across feature modalities, examines module-level disparities, and discusses operational constraints.

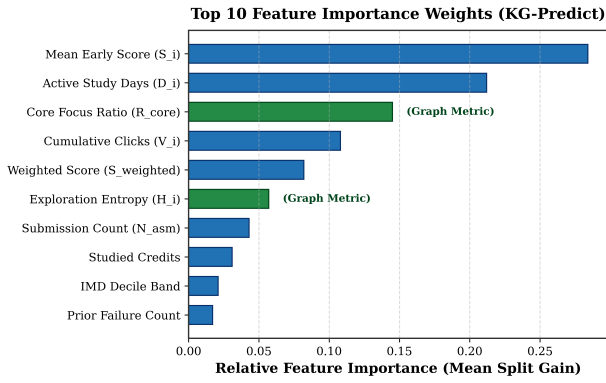


Figure 4: Top 10 feature importance weights (mean split gain) in the final KG-Predict classifier, highlighting graph metrics in green.

Table 4: Course-level performance breakdown.

Course module	Test samples	Accuracy (%)	F1-score (%)	AUC-ROC
CCC	924	78.35%	81.65%	0.8790
FFF	1,549	78.63%	78.15%	0.8693
EEE	570	78.25%	74.49%	0.8601
DDD	1,291	77.15%	80.00%	0.8548
BBB	1,537	76.51%	76.54%	0.8480
AAA	144	71.53%	59.41%	0.7370
GGG	504	66.47%	58.06%	0.6886

5.1. Modality-specific ablation and feature sensitivity

The progressive integration of feature modalities illuminates how distinct behavioral dimensions contribute to early student risk classification:

- Clickstream telemetry (Week 4 baseline): The demographic and scalar clickstream baseline established an AUC-ROC of 0.7902 and an accuracy of 71.25%. While raw click counts (V_i) and active days (D_i) serve as basic indicators of student presence, volume alone does not differentiate productive study from unfocused browsing.
- Graph-theoretic descriptors: Incorporating the Core-to-Passive Focus Ratio (R_{core}) and Node Exploration Entropy (H_i) improved classification precision from 75.14% to 80.33%. By penalizing superficial engagement (such as repeated visits to discussion boards without accessing core instructional content), the graph features reduced false-positive risk classifications.
- Early graded performance (Week 6 final): Integrating early assessment scores ($\bar{S}_i, S_{weighted}$) produced the largest performance gain, increasing the overall AUC-ROC to 0.8472 (+0.0570 gain) and F1-score to 76.99%. Feature importance ranking (Figure 4) confirmed that mean early assessment scores ($\bar{S}_i$) and active study days (D_i) carried the highest split gain weights in the final gradient-boosted ensemble.

5.2. Curricular Heterogeneity and Domain-Specific Disparities

A key empirical finding lies in the performance divergence observed across different academic disciplines in Table 4:

- STEM disciplines (CCC, FFF, EEE): The framework achieved its highest performance on STEM course offerings, with AUC-ROC scores ranging from 0.8601 to 0.8790 and accuracies above 78%. These curricula follow structured dependencies where early comprehension of foundational mathematical and programming concepts is prerequisite to subsequent exercises.
- Humanities and Social Sciences (AAA, GGG): In contrast, performance was lower on non-STEM modules, yielding AUC-ROC values of 0.7370 for AAA and 0.6886 for GGG. In humanities curricula, primary learning often occurs through offline textbook reading, critical essay composition, and independent reflection—activities that do not generate VLE interaction logs. This indicates that clickstream-driven graph modeling has inherent domain boundaries when applied to curricula that rely less heavily on continuous LMS activity.

5.3. Threats to validity and practical constraints

Two operational limitations should be considered:

1. Early cold-start sparsity: Prior to day 20, assessment submission records (A_i) are largely empty, limiting model utility during the initial two weeks of instruction.
2. Single-institution telemetry: While the OULAD corpus encompasses 32,593 enrollments across diverse cohorts, the data reflects a distance-learning institution; validation across synchronous, in-person university settings remains necessary.

6. Conclusion and future directions

This study presented KG-Predict, a multimodal predictive framework for early student risk identification in distance higher education. Addressing the limitations of scalar clickstream models, the framework constructs a directed Course Knowledge Graph directly from Learning Management System (LMS) relational tables and extracts closed-form topological descriptors (R_{core} and H_i) to capture navigational intent. Evaluated on 32,593 student registrations across 22 course offerings in the OULAD benchmark, the framework attained an AUC-ROC of 0.8472, an accuracy of 76.67%, and an F1-score of 76.99% at Week 6 of an academic term, representing improvements of +0.0570 in AUC-ROC and +5.42% in accuracy over baseline models.

Empirical evaluation across individual courses revealed that predictive capacity is discipline-dependent, delivering higher performance in structured STEM subjects (AUC-ROC up to 0.8790) while showing reduced efficacy in reading-intensive humanities courses (AUC-ROC of 0.6886).

Based on these findings, future research will pursue three concrete avenues:

1. Heterogeneous GNNs: Developing Relational Graph Convolutional Networks (R-GCNs) to learn continuous topological representations directly from relational databases, extending the discrete centrality and entropy metrics formulated in this work.
2. Cross-institutional transfer validation: Evaluating the generalizability of the framework on on-campus learning platforms (e.g., Canvas and Moodle) to quantify

performance under blended and in-person instructional models.

3. Actionable early-warning dashboard integration: Constructing localized explanation modules via TreeSHAP to provide academic advisors with student-specific risk attributions, translating predictive alerts into targeted instructional interventions before mid-term milestones.

Declarations and ethical statements

Conflict of interest: The authors declare that they have no conflicts of interest.

Funding statement: The authors declare that no specific funding was received for this research.

Artificial Intelligence usage statement: During the preparation of this manuscript, the authors utilized ChatGPT solely for language refinement and grammatical corrections. The authors carefully reviewed and revised the generated content and take full responsibility for the accuracy, integrity, and originality of the final manuscript.

Acknowledgment: The authors thank the Open University for providing open access to the OULAD benchmark dataset.

Availability of data and materials: The data and/or materials that support the findings of this study are available from the corresponding author upon reasonable request.

CRediT authorship contribution statement

Akshay Kumar: Conceptualization, Methodology & Writing—original draft, review and editing. **Neha Gulshan:** Conceptualization & Writing—review. **Varun Amrit:** Methodology, Writing—review.

Publisher's note

Krrish Scientific Publications Pvt. Ltd. and the *Journal of Applied Sciences and Modelling* remain neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- [1] Kuzilek J, Hlosta M, Zdrahal Z. Open University Learning Analytics dataset. *Scientific Data*. 2017 Nov 28;4(1):170171. Available from: <https://doi.org/10.1038/sdata.2017.171>
- [2] Adnan M, Habib A, Ashraf J, Mussadiq S, Raza AA, Abid M, Bashir M, Khan SU. Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models. *IEEE Access*. 2021 Jan 5;9:7519-39. Available from: <https://doi.org/10.1109/ACCESS.2021.3049446>
- [3] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016 Aug 13 (pp. 785-794). Available from: <https://doi.org/10.1145/2939672.2939785>
- [4] Albreiki B, Habuza T, Palakkal N, Zaki N. Clustering-based knowledge graphs and entity-relation representation improves the detection of at risk students. *Education and Information Technologies*. 2024 Apr;29(6):6791-820. Available from: <https://doi.org/10.1007/s10639-023-11938-8>
- [5] Niyogisubizo J, Liao L, Nziyumva E, Murwanashyaka E, Nshimyumukiza PC. Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*. 2022 Jan 1;3:100066. Available from: <https://doi.org/10.1016/j.caeai.2022.100066>
- [6] Lin Y, Chen H, Xia W, Lin F, Wang Z, Liu Y. A Comprehensive Survey on Deep Learning Techniques in Educational Data Mining: Y. Lin et al. *Data Science and Engineering*. 2025 Dec;10(4):564-90. Available from: <https://doi.org/10.1007/s41019-025-00303-z>
- [7] Almahdi AA, Sharef BT. Deep Learning Based An Optimized Predictive Academic Performance Approach. In *2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD)* 2023 Mar 8 (pp. 1-6). IEEE. Available from: <https://doi.org/10.1109/ITIKD56332.2023.10099652>
- [8] Xing W, Guo R, Petakovic E, Goggins S. Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Computers in human behavior*. 2015 Jun 1;47:168-81. Available from: <https://doi.org/10.1016/j.chb.2014.09.034>
- [9] Romero C, Ventura S. Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*. 2020 May;10(3):e1355. Available from: <https://doi.org/10.1002/widm.1355>
- [10] Baker RS. Artificial intelligence in education: Bringing it all together. *Digital education Outlook: Pushing the frontiers with AI, Blockchain, and Robots*. 2021 Jun 8:43-54. Available from: <https://books.google.co.in/books?id=yJ8yEAAAQBAJ&lpg=P443&ots=>
- [11] Hlosta M, Zdrahal Z, Zendulka J. Ouroboros: early identification of at-risk students without models based on legacy data. In *Proceedings of the seventh international learning analytics & knowledge conference* 2017 Mar 13 (pp. 6-15). Available from: <https://doi.org/10.1145/3027385.3029447>
- [12] Conijn R, Snijders C, Kleingeld A, Matzat U. Predicting Student Performance from LMS Data: A Comparison of 17 Blended Courses Using Moodle LMS. *IEEE Transactions on learning technologies*. 2016 Oct 12;10(1):17-29. Available from: <https://doi.org/10.1109/TLT.2016.2616312>
- [13] Xiao W, Ji P, Hu J. A survey on educational data mining methods used for predicting students' performance. *Engineering Reports*. 2022 May;4(5):e12482. Available from: <https://doi.org/10.1002/eng2.12482>
- [14] Wu Z, Pan S, Chen F, Long G, Zhang C, Yu PS. A Comprehensive Survey on Graph Neural Networks. *IEEE transactions on neural networks and learning systems*. 2020 Mar 24;32(1):4-24. Available from: <https://doi.org/10.1109/TNNLS.2020.2978386>
- [15] Karim-Abdallah B, Junior MA, Appiahene P, Harris E, Binful DK. Application of Machine Learning Algorithms in Predicting Academic Performance of Students in Higher Education Institutes (HEIs): A Systematic Review and Bibliographic Analysis. *African Journal of Applied Research*. 2025 Jan 11;11(1):536-59. Available from: <https://doi.org/10.26437/ajar.v11i1.869>
- [16] Xu L, Wang Z, Zhang S, Yuan X, Wang M, Chen E. Modeling Student Performance Using Feature Crosses Information for Knowledge Tracing. *IEEE Transactions on Learning Technologies*. 2024 Mar 22;17:1364-77. Available from: <https://doi.org/10.1109/TLT.2024.3381045>
- [17] Remegio AN. Predictive Analytics in Higher Education: A Systematic Literature Review of Machine Learning Models for Forecasting Academic Performance and Ethical Integration. In *2025 14th International Conference on Educational and Information Technology (ICEIT)* 2025 Mar 14 (pp. 213-218). IEEE. Available from: <https://doi.org/10.1109/ICEIT64364.2025.10976097>