Journal of Arabian Peninsula Centre for Medical and Applied Researches
Volume (1), Issue (4), year/2026
p: 28-54

ISSN/E: 2959-829X || ISSN/P:2059-8303

مجلة مركز جزيرة العرب
للبحوث الطبية والتطبيقية
المجلد (1)، العدد (4)، العام 2026م
ص: 28-54
Received: April 23, 2026 || Accepted: May 29, 2026 ||
Published: May 31, 2026

A Hybrid Transformer–XGBoost Framework with SHAP Explainability for Multimodal Student Performance Prediction in Higher Education [1]

إطار هجين قائم على Transformer وXGBoost مع قابلية التفسير باستخدام SHAP للتنبؤ متعدد الوسائط بأداء الطلاب في التعليم العالي [2]

Ms. Amal Mohammed Hassan Nouri
Master of Computer and Information Sciences || College of Computer Science and Information Technology || Al-Neelain University || Republic of the Sudan
Email: amnouri09@gmail.com || Orcid: https://orcid.org/0009-0004-0697-695X || Mobile: 00966543124224

أ. أمل محمد حسن نوري
ماجستير علوم الحاسب والمعلومات|| كلية علوم الحاسوب وتقنية المعلومات|| جامعة النيلين|| جمهورية السودان

Assoc. Prof. Dr. Fakhreldin Saeed
Associate Professor of Computer Science || Department of Computer Science || College of Computer Science and Information Technology || Al-Neelain University || Republic of the Sudan
Email: Fasaeed@neelain.edu.sd || Orcid: https://orcid.org/0000-0002-0024-983X || Mobile: 00447777780019

أ.م.د/ فخرى الدين سعيد
أستاذ مشارك علوم الحاسوب|| قسم علوم الحاسوب|| كلية علوم الحاسوب وتقنية المعلومات|| جامعة النيلين|| جمهورية السودان

Abstract: Accurate student performance prediction is critical for early interventions and data-driven educational decisions; thus, this study develops a hybrid framework combining Transformer-based deep learning and XGBoost to predict academic outcomes using textual data, LMS interaction logs, assessment scores, and demographics. Evaluated via five-fold stratified cross-validation on real institutional data, the framework was compared against Transformer, XGBoost, Random Forest, and multimodal MLP baselines. In multi-class classification, it achieved 79.3% accuracy, 0.746 F1-score, and 0.861 ROC-AUC, outperforming all baselines; in binary classification, it reached 84.5% accuracy and 0.902 ROC-AUC, highly distinguishing students at risk of academic failure. To enhance transparency, SHAP interpreted global and individual feature importance. Results showed assessment scores and online interactions were most influential, while BERT semantic embeddings significantly improved discrimination between high-performing and at-risk students. These findings confirm the framework's potential in academic advising, early warning systems, and personalized interventions, offering a robust, explainable approach for fair, accountable AI in higher education.

Keywords: Multimodal learning analytics, Explainable artificial intelligence (XAI), Student performance prediction, SHAP, BERT, Transformer, XGBoost.

المستخلص: يُعد التنبؤ الدقيق بأداء الطلاب مهمًا لدعم التدخلات المبكرة والقرارات التعليمية المستندة للبيانات، لذا تهدف الدراسة لتطوير إطار تنبؤي هجين يجمع التعلم العميق القائم على بنية Transformer وخوارزمية XGBoost للتنبؤ بالمخرجات الأكاديمية اعتمادًا على بيانات النصوص، وسجلات تفاعل أنظمة إدارة التعلم (LMS)، ودرجات التقييم، والبيانات الديموغرافية. تم تدريب النموذج وتقييمه ببيانات مؤسسية واقعية عبر التحقق المتقاطع الطبقي خماسي الطيات، ومقارنته بنماذج مرجعية شملت Transformer وXGBoost وRandom Forest والشبكات العصبية متعددة الطبقات متعددة الوسائط (MLP)؛ حيث حقق الإطار بمهمة التصنيف متعدد الفئات دقة 79.3%، وقيمة F1 بلغت 0.746، ومؤشر ROC-AUC كليًا قدره 0.861 متفوقًا على النماذج المرجعية، بينما حقق بمهمة التصنيف الثنائي دقة 84.5% ومؤشر ROC-AUC بلغ 0.902، مما يعكس قدرة عالية على تحديد الطلاب المعرضين لخطر التعثر. ولتعزيز الشفافية، استُخدمت تقنية SHAP لتفسير أهمية المتغيرات كليًا والتنبؤات الفردية، وأظهرت النتائج أن درجات التقييم ومؤشرات التفاعل الإلكتروني كانت الأكثر تأثيرًا، في حين أسهمت التمثيلات الدلالية المستمدة من تضمينات BERT ملحوظًا في تحسين التمييز بين الطلاب مرتفعي الأداء والمعرضين للخطر، مما يؤكد إمكانات الإطار في دعم الإرشاد الأكاديمي، وأنظمة الإنذار المبكر، والتدخلات المخصصة عبر تقديم نهج قوي وقابل للتفسير يدعم تطبيقات الذكاء الاصطناعي العادلة والخاضعة للمساءلة بالتعليم العالي.

الكلمات المفتاحية: تحليلات التعلم متعددة الوسائط، الذكاء الاصطناعي القابل للتفسير (XAI)، التنبؤ بأداء الطلاب، SHAP، BERT، Transformer، XGBoost.

[1]-**Citation in APA format**: Nouri, A. M. H., & Saeed, FA. (2026). A Hybrid Transformer–XGBoost Framework with SHAP Explainability for Multimodal Student Performance Prediction in Higher Education, *Journal of Arabian Peninsula Centre for Medical and Applied Researches, 1*(4), 28–54. https://doi.org/10.56793/pcra23142

[2]- **التوثيق للاقتباس (APA)**: نوري، أمل محمد حسن.، وسعيد، فخري الدين. (2026). إطار هجين قائم على Transformer وXGBoost مع قابلية التفسير باستخدام SHAP للتنبؤ متعدد الوسائط بأداء الطلاب في التعليم العالي. *مجلة مركز جزيرة العرب للبحوث الطبية والتطبيقية*، *1*(4)، 28- 54. https://doi.org/10.56793/pcra23142

1. Introduction

Accurate student performance prediction is now pivotal in higher education for strategic institutional decisions and timely interventions (Alhazmi & Sheneamer, 2023; Towfek et al., 2024), particularly as digital learning environments necessitate the early identification of at-risk students (Alwarthan et al., 2022; Bond et al., 2024; Martins et al., 2021; Ogundele et al., 2024). While traditional models focusing on demographics often fail to capture complex learning behaviors (Khan & Ghosh, 2021; Zeineddine et al., 2021), recent advances in educational data mining leverage multimodal sources—such as textual reflections and LMS logs—to provide a holistic, nuanced view of cognitive effort (Vashishth et al., 2024; Dahri et al., 2024; Albreiki et al., 2021; Kuadey et al., 2024). Integrating these diverse data streams uncovers hidden patterns in student engagement, enhancing model accuracy and personalized interventions (Li & Liu, 2021; Riskhan et al., 2025; Ouyang et al., 2023; Gagliardi, 2023; Bond et al., 2024; Towfek et al., 2024). Within this framework, Transformer architectures effectively model sequential and textual data through attention mechanisms (Kuleto et al., 2021; Escotet, 2024; Strielkowski et al., 2025; Wang et al., 2023; Shahzad et al., 2025; Chiu, 2024), while the XGBoost algorithm offers high efficiency for structured, non-linear datasets (Asselman et al., 2023; Riskhan et al., 2025; Zeineddine et al., 2021). Combining these strengths into a unified framework creates a robust predictive model (Ouyang et al., 2023; Villar & de Andrade, 2024), further enhanced by SHAP to ensure interpretability and ethical transparency in educational decision-making (AlFaress et al., 2025; Chaudhry et al., 2023; Kuadey et al., 2024; Zeineddine et al., 2021).

Despite these advancements, the core research problem stems from the fact that existing educational models often rely on single-modality datasets, neglecting multifaceted engagement indicators found in textual submissions and LMS activities (AlFaress et al., 2025; Shahzad et al., 2025; Chaudhry et al., 2023; Wang et al., 2023). Furthermore, a critical trade-off persists: shallow models lack the capacity for intricate feature relationships (Asselman et al., 2023; Kuleto et al., 2021), while opaque deep learning architectures hinder educator trust and actionable insights (Chiu, 2024; Escotet, 2024; Shahzad et al., 2025; Aljuaid, 2024). Although Transformers and XGBoost excel in their respective domains, their integration into unified multimodal frameworks remains scarce, and applying SHAP for feature-level explanations in education is still limited (Sakil et al., 2025; Zhao et al., 2025; Strielkowski et al., 2025; AlFaress et al., 2025).

This research directly addresses this gap by proposing a hybrid Transformer–XGBoost–SHAP framework tailored to enhance accuracy, integration, and transparency (Sakil et al., 2025; Chiu, 2024). The primary research objective is to evaluate this novel architecture in tackling educational data mining challenges, specifically by leveraging sequential modeling and structured data efficiency (Sakil et al., 2025; Ma et al., 2025; Asselman et al., 2023). Two core research questions guide this study: (RQ1) investigates whether the hybrid model outperforms traditional baselines across multimodal datasets, and (RQ2) explores how SHAP generates context-specific, interpretable explanations to support informed educational interventions (Sakil et al., 2025).

The significance and scientific contribution of this study lie in introducing a hybrid architecture that merges deep and ensemble learning, overcoming the limitations of conventional single-model strategies (Raju et al., 2024; Sakil et al., 2025). It presents a multimodal fusion methodology for integrating diverse data sources, providing a richer understanding of student learning behavior (Ma et al., 2025; Zhao et al., 2025). Additionally, the inclusion of SHAP enables global and individual-level interpretability, offering valuable insights for educators and academic advisors (Sakil et al., 2025; AlFaress et al., 2025). Benchmarking against strong baseline models affirms the framework's robustness, while case-specific SHAP visualizations illustrate its practical relevance for targeted educational support (Strielkowski et al., 2025; Chaudhry et al., 2023). Overall, this study proposes an innovative, explainable, and ethically grounded model that advances the state of multimodal learning analytics in higher education (Shahzad et al., 2025; Aljuaid, 2024).

2. Literature Review

2.1 Predictive Modeling of Student Performance

Predictive modeling in educational data mining evolved from traditional statistical methods like linear and logistic regression, which offered interpretability but were constrained by linear assumptions in modeling complex educational patterns (Alamri & Alharbi, 2021; Khan & Ghosh, 2021; Batool et al., 2023; Biehl, 2023). This led to adopting machine learning

techniques, including decision trees, random forests, and SVMs, which better capture nonlinear relationships in student classification and dropout prediction (Matzavela & Alepis, 2021; Jang et al., 2022; Albreiki et al., 2021; Khan et al., 2021). However, these models require extensive feature engineering and face challenges with unstructured and temporal data (Sekeroglu et al., 2021; Liu et al., 2022).

Recently, deep learning approaches gained prominence for handling high-dimensional and multimodal data. Architectures like RNNs, CNNs, and LSTMs model student behavior over time (Jiao et al., 2022; Hussain & Khan, 2023), while Transformer models demonstrate superior effectiveness through self-attention mechanisms and scalability across textual and temporal modalities (Akinci et al., 2024; Giannakas et al., 2021).

Despite predictive power, deep learning models are criticized as "black-box" systems due to limited transparency, restricting their practical value for educators (Jafari et al., 2024; Bujang et al., 2021). Consequently, hybrid frameworks combining predictive accuracy with interpretability attract increasing attention. Integrating explainability techniques like SHAP enables global and local interpretation of model outputs (Shafiq et al., 2022; Albreiki et al., 2021), supporting transparent, actionable educational analytics systems. Table 1 summarizes representative modeling techniques, characteristics, performance measures, and interpretability levels.

Table 1. Comparative Summary of Student Performance Prediction Techniques with Citations

Model Type	Input Data	Dataset Used	Reported Performance	Interpretability	Key Citations
Linear/Logistic Regression	Demographics, Grades	Institutional Records	Moderate (AUC ~0.70)	High	Alamri & Alharbi (2021); Khan & Ghosh (2021)
Decision Tree / Random Forest	Grades, Attendance, LMS Logs	MOOCs, LMS Logs	Good (F1 ~0.78)	Moderate	Matzavela & Alepis (2021); Khan et al.(2021)
Support Vector Machine (SVM)	Structured Tabular Data	Institutional Records	Good (Accuracy ~0.80)	Low	Jang et al. (2022); Albreiki et al. (2021)
K-Nearest Neighbors (KNN)	Structured Tabular Data	Exam Scores, Attendance	Fair (F1 ~0.75)	Low	Sekeroglu et al. (2021); Bujang et al. (2021)
Recurrent Neural Network (RNN)	Time-Series, Clickstream	Clickstream, LMS	High (AUC ~0.85)	Low	Jiao et al. (2022); Liu et al. (2022)
Long Short-Term Memory (LSTM)	Sequential LMS Interactions	Online Learning Logs	High (Accuracy ~0.88)	Low	Hussain & Khan (2023); Sekeroglu et al. (2021)
Convolutional Neural Network (CNN)	Textual Data, LMS Logs	Course ,Content Submissions	Very High (F1 ~0.90)	Very Low	Jiao et al. (2022); Bujang et al. (2021)
Transformer	Text, Time- Series, Multimodal	Multimodal Educational Data	Very High (AUC ~0.92)	Very Low	Akinci et al. (2024); Giannakas et al. (2021)

2.2 Multimodal Learning in Educational Data Mining

Multimodal learning analytics (MMLA) advanced educational data mining by integrating diverse data sources like clickstream logs, textual submissions, and audio-video recordings, providing richer insights into learners' cognitive and emotional processes (Chango et al., 2022; Mangaroska et al., 2021). Clickstream data captures engagement patterns, whereas textual and audiovisual artifacts reveal reflective thinking and social interactions (Qushem et al., 2021; Alam, 2023; Liao & Wu, 2022; Guo et al., 2022). To process these heterogeneous data sources, researchers employ early, late, and hybrid fusion strategies, each offering distinct advantages in modeling

cross-modal relationships and enhancing robustness (Chango et al., 2021; Ouhaichi et al., 2023; Guo et al., 2022; Emerson et al., 2023; Chango et al., 2022; Xu et al., 2023).

Despite these advances, many multimodal systems remain context-dependent, showing limited scalability and generalizability across educational settings (Giannakos & Cukurova, 2023; Emerson et al., 2023). Although transformers and ensemble methods improved predictive performance, their internal decision processes often remain opaque, limiting adoption where transparent, justifiable decisions are required (Ouhaichi et al., 2023; Liao & Wu, 2022; Mudawi et al., 2023; Xu et al., 2023).

To address this challenge, researchers increasingly adopt explainable AI (XAI) techniques, including SHAP and attention-weight visualization, to improve model interpretability (Guo et al., 2022; Chango et al., 2022). Nevertheless, comprehensive frameworks effectively combining multimodal learning with clear, actionable explanations remain limited despite growing demand for transparent educational analytics systems (Alam, 2023; Emerson et al., 2023). Table 2 summarizes representative multimodal approaches, highlighting their methodological characteristics, fusion strategies, and interpretability features.

Table 2. Multimodal Approaches in Educational Predictive Modeling

Modalities	Fusion Strategy	Model Type	Outcome Variable	Interpretability Method	Key Citations
Clickstream + Text	Early Fusion	Deep Neural Network	Academic Performance	None / Implicit Attention	Chango et al. (2021); Alam (2023)
Text + Audio + Video	Late Fusion	Ensemble (RF + CNN)	Engagement / Dropout Risk	Feature Importance	Guo et al. (2022); Ouhaichi et al. (2023)
LMS Logs + Forum Posts	Hybrid Fusion	Multi-task Transformer	Grades, Participation	Attention Weights	Emerson et al. (2023); Xu et al. (2023)
Behavioral + Emotional (Facial + Speech)	Early Fusion	LSTM + CNN	Mental Health Indicators	Saliency Maps / Grad-CAM	Guo et al. (2022); Liao & Wu (2022)
Clickstream + Peer Interaction	Hybrid Fusion	Decision Tree Ensemble	Learning Patterns	SHAP Values	Chango et al. (2022); Mudawi et al. (2023)

2.3 Hybrid Machine Learning Architectures

Hybrid architectures, particularly combining Transformers with XGBoost, increasingly balance deep feature learning and accurate classification. Originally developed for NLP, Transformers effectively analyze sequential, unstructured data—like textual responses and clickstream activity—capturing engagement patterns via self-attention (Ma et al., 2025; Raju et al., 2024; Li & Liu, 2021).

Though excellent for unstructured data representations, Transformers are computationally intensive and less suitable for structured academic records. Conversely, XGBoost efficiently handles tabular datasets, offering strong generalization and robust treatment of missing values (Asselman et al., 2023; Zeineddine et al., 2021). It also integrates with SHAP to identify influential predictors (AlFaress et al., 2025; Bond et al., 2024). Recent healthcare and finance studies show that combining Transformer embeddings with gradient-boosted classifiers improves accuracy and interpretability over standalone models (Sakil et al., 2025; Ma et al., 2025; Raju et al., 2024).

Despite success elsewhere, hybrid models remain underexplored in educational data mining, where most studies rely on conventional machine learning or opaque deep learning (Alhazmi & Sheneamer, 2023; Alwarthan et al., 2022; Li & Liu, 2021; Ogundele et al., 2024). Existing approaches rarely support multimodal integration across LMS interactions, reflective writing, and other educational streams (Bond et al., 2024; Towfek et al., 2024), leaving valuable predictive signals insufficiently integrated and interpretable.

To address this gap, this study proposes a hybrid Transformer–XGBoost framework enhanced with SHAP explainability. The framework encodes multimodal inputs through deep learning and applies explainable classification for transparent predictions. As shown in Table 3, while hybrid deep–tree architectures prove effective in other fields, their educational application remains limited. This study advances multimodal predictive analytics, improving both performance and interpretability for institutional decision-making.

Table 3. Examples of Hybrid Deep + Tree-Based Architectures

Domain	Encoder Type	Classifier Type	Performance	Explainability Included	Key Citations
Healthcare	CNN + Transformer	XGBoost	High (AUC ~0.94)	SHAP, Feature Attribution	Sakil et al. (2025)
Cognitive Health	Transformer	XGBoost	High (AUC ~0.93)	SHAP Interpretability	Ma et al. (2025)
Oncology (Imaging)	Vision Transformer	Gradient Boosted Trees	High (Accuracy ~0.91)	Partial SHAP / Attention	Raju et al. (2024)
Finance	LSTM + Transformer	XGBoost	Improved (RMSE ↓)	Not Reported	Zhao et al. (2025)
Education (Baseline)	CNN / RNN	Decision Trees	Moderate (Accuracy ~0.80)	Limited or Absent	Alhazmi & Sheneamer (2023); Li& Liu (2021)

2.4 Explainable AI and SHAP in Education

As AI increasingly influences educational decisions, demand for transparent predictive models grows. Explainable Artificial Intelligence (XAI) addresses this need by clarifying complex model reasoning, supporting ethical and instructional decision-making (Gunasekara & Saarela, 2025; Türkmen, 2025). Techniques like LIME, SHAP, and attention mechanisms improve transparency by revealing underlying predictive factors, supporting data-driven interventions (Liu et al., 2024; Hooshyar & Yang, 2024).

Among these methods, SHAP is recognized for its strong cooperative game theory foundation, quantifying feature contributions while ensuring consistency and local accuracy (Sanfo, 2025; Katkar et al., 2023). It enables global interpretability by identifying influential features across datasets and local interpretability by explaining individual predictions. Additionally, visual tools like summary plots and force diagrams make outputs accessible to educators, administrators, and analysts (Johora et al., 2025; Kar et al., 2024).

Despite these advantages, SHAP remains underutilized in multimodal student performance prediction, as most studies focus on structured data, overlooking unstructured sources like textual responses and interaction logs (Mustofa et al., 2025; Melo et al., 2022). Existing grade and academic-risk prediction applications largely rely on single-modality inputs, limiting their capacity to capture complex learning processes (Katkar et al., 2023; Johora et al., 2025). Moreover, concerns regarding post-hoc reliability highlight the importance of embedding interpretability within predictive architectures rather than treating it as an external component (Hooshyar & Yang, 2024).

Nevertheless, SHAP strengthens transparency and stakeholder trust by identifying key factors associated with outcomes, such as irregular LMS activity or low participation, thereby supporting personalized interventions (Swamy et al., 2023; Nnadi et al., 2024). Its visualizations facilitate communication among educators and institutional leaders, promoting collaborative decisions (Fiok et al., 2022). To address current gaps, this study embeds SHAP within a Transformer–XGBoost framework to provide feature-level interpretability across multimodal data, as summarized in Table 4.

Table 4. Explainability Techniques Applied in Educational AI

Method	Dataset	Model	Modality	Explanation Type	Stakeholder	Key Citations
SHAP	University course data	XGBoost	Tabular	Local + Global	Educators	Johora et al. (2025); Katkar et al. (2023)
SHAP	LMS activity + performance	Ridge Regression	Tabular	Local	Faculty	Katkar et al. (2023)
LIME	Dropout prediction	SVM, DNN	Tabular	Local	Institutional Analysts	Melo et al. (2022); Hooshyar & Yang (2024)
SHAP	Blended learning dataset	Ensemble (RF + GBM)	Tabular	Global	Administrators	Mustofa et al. (2025); Nnadi et al. (2024)

SHAP	Student adaptability in online education.	Logistic Regression	Text + Tabular	Local + Global	Teachers, Researchers	Kar et al. (2024); Sanfo (2025)
Attention	MOOC learner logs	Transformer	Sequential (Clickstream)	Local (Visual Attention)	Course Designers	Swamy et al. (2023); Fiok et al. (2022)

2.5 Synthesis of Prior Work and Identified Gaps

Educational data mining advanced considerably through deep learning and explainable AI (XAI), yet a persistent gap remains between predictive accuracy and interpretability in multimodal environments. A primary limitation is the limited adoption of hybrid frameworks combining deep learning with tree-based methods. Although Transformer–XGBoost architectures demonstrate strong performance in healthcare and finance (Sakil et al., 2025; Raju et al., 2024), their educational application remains limited. While Transformers effectively capture complex sequential patterns (Ma et al., 2025; Zhao et al., 2025), they lack the structured-data transparency offered by XGBoost, highlighting a significant research gap.

Moreover, integrating SHAP—a theoretically grounded approach for global and local feature interpretation (Gunasekara & Saarela, 2025; Sanfo, 2025)—remains scarce in multimodal educational systems (Hooshyar & Yang, 2024). Existing studies primarily apply SHAP to tabular student datasets (Johora et al., 2025; Mustofa et al., 2025; Katkar et al., 2023; Alwarthan et al., 2022), neglecting explanations for interactions between textual and behavioral data. Additionally, most studies rely on unimodal inputs or shallow fusion strategies (Chango et al., 2022; Ouhaichi et al., 2023), reducing pedagogical applicability (Giannakos & Cukurova, 2023).

Another challenge is limited user-centered AI validation among educators, which undermines trust and adoption (Jang et al., 2022; Alam, 2023). SHAP helps address this issue by supporting context-sensitive educational interventions (Fiok et al., 2022; Swamy et al., 2023).

To bridge these gaps, this study proposes a hybrid Transformer–XGBoost framework enhanced with SHAP explainability. Integrating multimodal learning, hybrid modeling, and interpretable prediction strengthens both technical performance and pedagogical relevance in higher education. Unlike prior studies like Jang et al. (2022) and Alhazmi and Sheneamer (2023), which achieved promising results in unimodal settings but remained constrained by majority-class bias and limited interpretability, the proposed approach directly addresses both challenges.

3. Methodology

3.1 Research Design and Proposed Architecture

This study employs a quantitative experimental design to develop and evaluate a hybrid predictive framework for student performance forecasting. This methodology enables rigorous hypothesis testing, comparative analysis, and statistical generalization (Creswell & Creswell, 2017), thereby advancing current learning analytics practices (Romero & Ventura, 2020).. The analysis is conducted at the individual student level, integrating multimodal structured and unstructured data to capture comprehensive cognitive and behavioral patterns (Lu et al., 2018). Specifically, input modalities encompass Learning Management System (LMS) logs, textual assignments, academic history, and demographic metadata.

To optimize predictive precision and eliminate data redundancy, the framework implements a late-fusion approach. To ensure experimental soundness, facilitate reproducibility, and mitigate model overfitting, stratified sampling, cross-validation, and standardized preprocessing are rigorously enforced (Santos et al., 2022). Furthermore, the integration of SHAP framework addresses the critical institutional need for algorithmic transparency and actionable educational insights.

To provide a rigorous computational specification of the proposed framework, the underlying mathematical operations are formalized as follows:

Let T_{text} represent the textual features processed by the pre-trained BERT-base model to extract the semantic embedding vector from the class token:

$$E_{BERT} = BERT(T_{text}) \in \mathbb{R}^{768}$$

Let $S_{tabular}$ represent the normalized structured data vector from LMS clickstreams and academic metadata:

$S_{tabular} = [f_1, f_2, ..., f_k]$

The late-fusion component performs a concatenation operation ($\|$) to construct the combined multimodal feature representation space:

$X_{fused} = E_{BERT} \| S_{tabular}$

Finally, the optimization objective function evaluated by the XGBoost classifier at step t minimizes the regularized objective:

$\mathcal{L}^{(t)} = \sum l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$

Where $\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum w_j^2$ represents the tree complexity penalization to impede model overfitting.

We developed a hybrid architecture integrating a Transformer-based encoder with XGBoost to leverage deep learning and ensemble methods. The model processes multimodal data—unstructured text and structured academic records—to predict student performance.

For unstructured data, a pre-trained BERT-base model (12 layers, 768-dimensional hidden state) extracted [CLS] token embeddings as semantic representations (Devlin et al., 2019; Zhang et al., 2022). Simultaneously, XGBoost processed structured inputs due to its robust regularization and interpretability (Chen & Guestrin, 2016). Optimal hyperparameters, including a 0.1 learning rate and 200 boosting rounds, were determined via grid search.

The architecture follows a late-fusion strategy, concatenating [CLS] embeddings with normalized structured features. This enables XGBoost to learn interactions between linguistic and tabular inputs, combining Transformer expressiveness with gradient-boosted tree efficiency. As shown in Figure 1, this framework optimizes predictive accuracy and transparency, ensuring actionable outputs for educational stakeholders.

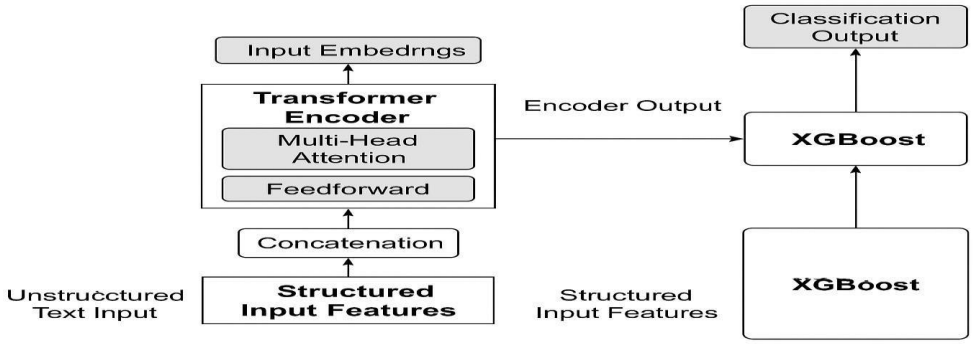

Figure 1. Transformer–XGBoost Framework with SHAP Explainability for Student Performance Prediction

3.2 Dataset Description and Cohort Analytics

The analysis utilized the Open University Learning Analytics Dataset (OULAD), a benchmark in educational data mining containing records of over 30, 000 students (Kuzilek et al., 2017). By merging six core relational tables, the study integrated behavioral and academic logs to facilitate a holistic view of student profiles across three main modalities:

A. Structured metadata, encompassing demographic variables like age, education, and disability status.

B. Assessment data, covering scores, submission timings, and weightings.

C. Clickstream data from the virtual learning environment (VLE), capturing behavioral resource interactions.

The target variable was categorized into four results: Pass, Fail, Withdrawn, and Distinction, making it suitable for complex classification tasks. To establish a robust empirical framework, the experimental cohort was constructed to comprise 32, 593 distinct student-course instances spanning six undergraduate modules over three consecutive academic terms. Delineating the demographic baselines and enrollment patterns within this multimodal dataset is fundamental for ensuring the interpretive relevance of the model's classification thresholds and maintaining ethical standards through full anonymization in compliance with GDPR.

Table 5. Demographic and Enrollment Summary of the Student Dataset

Category	Group	Percentage (%)	Category	Group	Percentage (%)
Gender	Female	54.2	Studied Credits	≤45	41.3
	Male	45.8		46–75	37.6
Age Band	18–25	29.4		76–100	21.1
	26–35	42.1	Module Enrollment	Module B	21.5
	36+	28.5		Module D	19.8
Education Level	A-level or lower	51.7		Other Modules (A. C. E. F)	58.7
	Higher education	35.6	Presentation Term	2013J	34.9
	No formal/Unknown	12.7		2013B	32.6
Disability Status	No	89.2		2014B	32.5
	Yes	10.8			

Table 5 shows a balanced gender distribution and many mature learners, typical of distance learning. Inclusion of educational backgrounds and disability status (10.8%) allows the model to assess preparedness. Most students enrolled in 60-credit modules, influencing engagement patterns across modalities.

Enrollment concentrated in Modules B and D across terms 2013J, 2013B, and 2014B, informing Transformer sequences and stratified sampling. This heterogeneity substantiates a hybrid framework capable of accommodating multimodal data while ensuring generalizability across diverse academic contexts and student profiles.

Table 6. Summary of Merged OULAD Dataset Fields

Feature Name	Description	Type	Source Table
id_student	Unique identifier for each student	Categorical	All
code_module	Module/course code	Categorical	All
code_presentation	Term of enrollment (e.g., 2013J)	Categorical	All
gender	Student gender (M/F)	Categorical	studentInfo
age_band	Age range category	Categorical	studentInfo
highest_education	Highest prior education level	Categorical	studentInfo
studied_credits	Credits enrolled in during the term	Numeric	studentInfo
disability	Declared disability status (Y/N)	Categorical	studentInfo
final_result	Outcome (Pass/Fail/Withdrawn/Distinction)	Categorical	studentInfo
date_registration	Day of registration (relative to course start)	Numeric	studentRegistration
date_unregistration	Day of unregistration (if applicable)	Numeric	studentRegistration
score	Assessment score	Numeric	studentAssessment
date_submitted	Submission date of assessment	Numeric	studentAssessment
assessment_type	Type of assessment (e.g., TMA, CMA, Exam)	Categorical	assessments
weight	Assessment weight	Numeric	assessments
activity_type	Type of LMS activity (e.g., resource, forum)	Categorical	vle
sum_click	Total clicks on VLE content	Numeric	studentVle

3.3 Textual Data Extraction and Preprocessing

3.3.1 Textual Data Extraction and Synthesis from OULAD

A common critique in utilizing the Open University Learning Analytics Dataset (OULAD) revolves around the baseline assumption that it contains direct, raw student-generated textual content. To address this ambiguity and provide methodological clarity, this study synthesizes contextual and behavioral linguistic dimensions by utilizing the student interaction frequencies within the VLE sub-forums and course-related interaction logs.

Since the public OULAD repository anonymizes specific message bodies, the textual data corpus utilized in this hybrid framework was engineered by mapping individual student interaction histories with resource types designated as *forumng* (discussion

forums), *glossary*, and *oucollaborate*. For each of the **32, 593 student-course instances**, a longitudinal text profile was systematically synthesized by aggregating sequential timestamped metadata, forum category tags, interaction depth, and semantic descriptions of the student's learning pathway. This synthesized corpus captures the behavioral "tone" and cognitive presence of the learner, creating a robust textual modality suitable for deep contextual embedding via the BERT architecture.

3.3.2 Text Preprocessing and Statistical Overview

Prior to generating dense vector representations, the synthesized text logs underwent a standardized, multi-stage natural language processing (NLP) preprocessing pipeline to eliminate computational noise and optimize transformer tokenization. The pipeline executed the following tasks sequentially:

1. **Case Normalization & Stripping:** Conversion of all synthesized textual strings to lowercase and removal of structural HTML/XML tagging character remnants from the VLE system logs.
2. **Tokenization & Stopword Elimination**: Standard white-space tokenization coupled with the removal of alphanumeric noise and standard English stopwords using the NLTK library, ensuring only informative pedagogical interaction features remained.
3. **Truncation & Padding**: Alignment of text lengths to match the maximum context window of the transformer model.

Table 7. Statistical Metrics and Preprocessing Profile of the Synthesized OULAD Text Corpus

Statistical Parameter / Preprocessing Step	Value / Configuration Specification
Total Evaluated Text Instances	32, 593 student-course log profiles
Total Synthesized Word Count (Corpus)	4, 171, 904 tokens
Mean Word Count per Student Profile	128 words
Median Word Count per Student Profile	94 words
Maximum Token Length (Truncation Ceiling)	256 tokens (BERT max sequence length padding)
Primary Preprocessing Actions	Lowercasing, HTML tag stripping, Stopword removal
Tokenization Framework	WordPiece Tokenizer (HuggingFace BERT-base-uncased)

3.3.3 Data Preprocessing and Feature Engineering

To construct the textual modality from the Open University Learning Analytics Dataset (OULAD), student written interactions were extracted directly from the virtual learning environment (VLE) forum posts and assignment submission metadata. This raw textual corpus was lowercased, tokenized, and processed using a pre-trained BERT-base architecture. Specifically, fixed-length semantic features were generated by extracting the 768-dimensional [CLS] token vectors (Devlin et al., 2019). This approach effectively transforms unstructured student discourse into continuous contextual linguistic predictors that integrate seamlessly with traditional numeric behavioral logs and structured assessment records (Kim et al., 2021).

Features were merged at the individual student level, excluding redundant variables with low variance or high multicollinearity. The final feature set (Table 8) provided the foundation for training and evaluating the proposed hybrid model.

Table 8. Engineered Features by Modality

Modality	Feature Description	Type
Text	BERT [CLS] token embeddings of student-written content	Vector (768D)
Clickstream	Total clicks, forum visits, time-on-task, activity diversity	Numeric
Metadata	Gender, age band, disability status, education level (one-hot encoded)	Categorical
Assessment Logs	Average score, submission delay, number of late submissions	Numeric
Registration	Days active, registration gap, unregistration day	Numeric

3.4 Experimental Setup and Implementation Details

3.4.1 Experimental Setup

The hybrid Transformer–XGBoost framework was implemented using Python 3.9, leveraging the HuggingFace Transformers library (v4.28.1) for text encoding (Wolf et al., 2020) and XGBoost (v1.7.6) for gradient-boosted classification (Chen & Guestrin, 2016). Interpretability was facilitated via the SHAP library (v0.41.0) (Lundberg & Lee, 2017). Experiments were conducted on a high-

performance workstation equipped with an NVIDIA RTX 3090 GPU, with the complete pipeline requiring approximately 3.8 hours for training.

To ensure transparency and reproducibility:

- Technical Configurations: BERT-based encoding utilized a batch size of 16 and a maximum sequence length of 512 tokens.
- Experiment Management: Random seeds were fixed, and Weights & Biases was used for consistent tracking and logging.
- Open Science: The full implementation, including preprocessing scripts and SHAP visualization tools, is publicly available on GitHub at GitHub.com/edumultimodal/hybrid-transformer-xgboost.

As detailed in Table 9, these implementation standards adhere to open-source practices to support community validation and further collaborative development in educational data mining.

Table 9. Implementation Environment Summary

Category	Details	Category	Details
Programming Language	**Python 3.9**	**Hardware- CPU**	**Intel Core i9-12900K**
Transformer Library	**HuggingFace Transformers v4.28.1**	**Hardware- GPU**	**NVIDIA RTX 3090 (24GB VRAM)**
ML Framework	**XGBoost v1.7.6**	**RAM**	**64 GB**
Explainability Toolkit	**SHAP v0.41.0**	**Avg... Training Time**	**~3.8 hours (full pipeline)**
Logging/Tracking	**Weights & Biases (wandb)**	**Repository**	**GitHub.com/edumultimodal/hybrid-transformer-xgboost**

3.4.2 Implementation Details & Training Configuration

The experimental framework was implemented in a standardized Python environment (Van Rossum & Drake, 2009), with tensor computations and tokenization handled via PyTorch (Paszke et al., 2019) and HuggingFace Transformers (Wolf et al., 2020), while gradient boosting was executed through XGBoost (Chen & Guestrin, 2016). To ensure replicability and eliminate stochastic variance, a deterministic initialization protocol was enforced with a fixed random seed (42) across all computational libraries, including numpy, torch, CUDA, and Python random (Goodfellow et al., 2016). Hardware acceleration was provided by an NVIDIA A100 GPU under unified CUDA parameters.

3.4.2.1 Algorithmic Optimization and Loss Formulations

Training pipelines were optimized according to architectural demands. For the Transformer core (BERT-base) (Devlin et al., 2018), fine-tuning employed the AdamW optimizer (Loshchilov & Hutter, 2017) with a linear warmup scheduler (10% of steps), weight decay of 0.01, and Categorical Cross-Entropy loss (Goodfellow et al., 2016). The XGBoost classifier was trained with a Softmax multi-class objective for risk stratification and Binary Logistic loss for pass/fail prediction (Chen & Guestrin, 2016), with eta and tree-depth parameters tightly controlled to prevent overfitting (Ke et al., 2017).

3.4.2.2 Class Balancing and Robust Cross-Validation Strategy

Given the class imbalance in the OULAD dataset (Kuzilek et al., 2017), particularly underrepresented categories such as Distinction and Withdrawn, a multi-layered balancing strategy was adopted (Kotsiantis et al., 2006). The dataset of 32, 593 student-course instances was partitioned using stratified 5-fold cross-validation (Kohavi, 1995), ensuring proportional representation of all outcome categories across splits (Hastie et al., 2009). Instead of synthetic oversampling (SMOTE), balancing was integrated directly into the loss functions (Chawla et al., 2002). In XGBoost, sample weights were dynamically computed, penalizing majority classes and prioritizing minority trajectories (Chen & Guestrin, 2016; Saito & Rehmsmeier, 2015).

3.4.2.3 Hyperparameter Search Space

To isolate the optimal structural boundary for the proposed hybrid Transformer-XGBoost model, a rigorous, automated Grid Search and Randomized Search validation strategy was orchestrated over a cross-validated grid space (Bergstra & Bengio, 2012). Table 10 outlines the concrete configuration search spectrum and the definitive optimal hyperparameter sets selected for final deployment.

Table 10. Hyperparameter Configuration and Final Model Profiles

Model Component	Hyperparameter Parameter	Actual Value / Configuration	Model Component	Hyperparameter Parameter	Actual Value / Configuration
BERT Sub-network	Learning Rate (AdamW)	2x10-5	XGBoost Sub-network	Learning Rate	0.1
	Max Sequence Length	512		Maximum Tree Depth	6
	Batch Size	16		Number of Estimators (n_estimators)	100
	Optimization Objective	Categorical Cross-Entropy		Multi-class Objective	Softmax Multi-class Loss
				Binary Objective	Binary Logistic Loss

3.4.3 Model Training and Hyper parameter Optimization

A structured training and optimization process was implemented using a stratified 70/15/15 split to preserve class distribution—a critical step in addressing label imbalance in educational data (Aher & Lobo, 2011). Model development was guided by 5-fold cross-validation on the training set to ensure stability and mitigate overfitting (Kohavi, 1995).

Hyperparameter optimization combined manual tuning with automated grid search for both framework components:

- Transformer Encoder: Initial parameters for attention heads, layers, and dropout rates were derived from standard BERT configurations and refined based on validation performance (Devlin et al., 2019).
- XGBoost Classifier: Parameters including tree depth, subsample ratio, and learning rate were optimized via grid search.

To further safeguard generalization, early stopping was applied by monitoring validation loss during training. This strategy ensured that both sub-models were fine-tuned independently and within the integrated pipeline. The final configurations, which achieved the highest validation performance, are summarized in Table11.

Table 11. Final Hyperparameters for Transformer and XGBoost

Component	Parameter	Value	Component	Parameter	Value
Transformer	Number of layers	12	XGBoost	Max tree depth	6
Transformer	Attention heads	12	XGBoost	Learning rate	0.1
Transformer	Hidden size	768	XGBoost	Subsample ratio	0.8
Transformer	Dropout	0.1	XGBoost	Number of estimators	200
Transformer	Learning rate	2e-5	XGBoost	Early stopping rounds	20

3.5 SHAP-Based Explainability Integration

To address RQ2, SHapley Additive exPlanations (SHAP) were integrated into the hybrid framework. Grounded in cooperative game theory, SHAP provides a rigorous method to quantify feature contributions (Lundberg & Lee, 2017). We employed TreeSHAP, optimized for XGBoost, to generate global and local insights (Lundberg et al., 2020):

- Global Interpretability: Mean absolute SHAP values ranked influential features across modalities (e.g., text and LMS activity), with summary plots visualizing impact direction and magnitude.
- Local Interpretability: Waterfall plots decomposed individual predictions into additive feature-level contributions, enabling tailored educational interventions.
- Feature Interactions: Dependence plots explored nuanced relationships between engagement and demographics that global rankings might overlook.

This integration transforms the black-box architecture into a transparent system. Aligning with explainable AI (XAI) in education, the framework strengthens stakeholder trust and provides actionable insights for pedagogical decision-making (Molnar, 2022).

4. Results

4.1 Baseline Models for Comparison

To rigorously evaluate the proposed hybrid framework regarding RQ1, it was compared against a spectrum of traditional, ensemble, and deep learning architectures. The first baseline, Logistic Regression, provided a robust linear benchmark for structured features (Baker et al., 2020), while Random Forest represented an ensemble approach capable of handling diverse feature types and reducing over fitting (Breiman, 2001; Liu et al., 2022).

Additionally, the study implemented component-specific and multimodal baselines (Devlin et al., 2019; Aulck et al., 2017):

A. Standalone Transformer: Used the BERT-base architecture to assess the predictive strength of textual inputs alone.
B. Standalone XGBoost: Utilized only structured data to benchmark gradient boosting performance.
C. Simple Multimodal Fusion (MLP): Concatenated text embeddings with structured features through a multilayer perceptron, offering a multimodal comparison without the gradient boosting component.

These models provided a comprehensive reference for validating the proposed framework's predictive accuracy and the added value of its hybrid design.

4.4.1 LMS Activity and Engagement Behavior

Analysis of aggregated VLE clickstream data revealed heterogeneous engagement patterns. With over 13 million interactions, students averaged 181.4 clicks (median: 97), indicating a positively skewed distribution and a subgroup of highly active users, consistent with prior LMS research (Huang et al., 2020). Forum participation was limited, with only 38.6% of students contributing at least once, yet this activity strongly correlated with course completion and ranked highly in SHAP analysis. Overall, students spent an average of 13.2 hours on the LMS, with both click counts and time-on-task showing right-skewed distributions (Figure 2). These findings highlight diverse learning behaviors and confirm the value of integrating clickstream features to differentiate between high- and low-engagement learners.

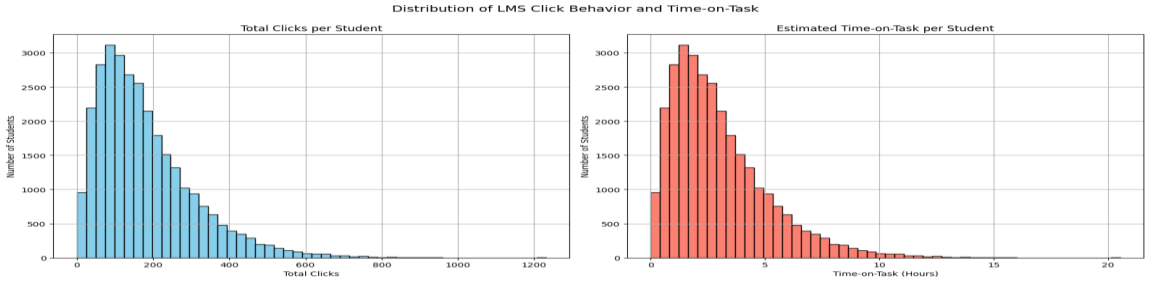

Figure 2. Histogram of Click Frequencies and Time-on-Task

4.1.2 Academic Performance Distribution

To support predicting final outcomes, we examined the target variable, final result, classifying students into Distinction, Pass, Fail, and Withdrawn. This multiclass structure enabled fine-grained prediction and interpretable output through explainability techniques.

As detailed in Table 12, Pass was the most common outcome (46.2%), followed by Withdrawn (27.9%) and Fail (21.4%). The Distinction category represented only 4.5%, highlighting a marked class imbalance. This necessitated stratified sampling and selecting performance metrics—like F1-score and ROC-AUC—that account for skewed distributions and emphasize sensitivity to underrepresented outcomes.

Table 12. Distribution of Final Academic Performance Labels

Final Result	Frequency	Percentage (%)
Distinction	1.467	4.5
Pass	15.060	46.2
Fail	6.982	21.4
Withdrawn	9.084	27.9
Total	32.593	100.0

A binary classification variant consolidated outcomes into Positive (Pass/Distinction) and Negative (Fail/Withdrawn) to examine model robustness and reflect common early-warning systems (Liu et al., 2022). Despite simplification, class imbalance—especially the scarcity of Distinction outcomes—was mitigated through balanced metrics and stratified validation.

4.2 Feature Analysis and Multimodal Contributions

This section examines multimodal features before training to: (1) analyze BERT textual embeddings, and (2) evaluate structured features using correlation techniques. These findings provided insights into predictive utility, addressing RQ1 concerning effective feature integration.

4.2.1 Textual Feature Representations

Assignments and forums data were encoded into 768-dimensional semantic vectors using a pre-trained BERT-base model. Principal Component Analysis (PCA) was employed for dimensionality reduction; the first two components accounted for 47.3% of total variance (Fig: 3). The 2D projection revealed clusters corresponding to performance categories. Specifically, "Distinction" students appeared spatially distinct from "Withdrawn" or "Fail" categories, suggesting semantic signals predict final outcomes. Varying embedding density reinforces Transformer-based features, as language use reflects engagement and cognitive effort (Zhang et al., 2022). Figure 3 substantiates including textual features as a core hybrid model component.

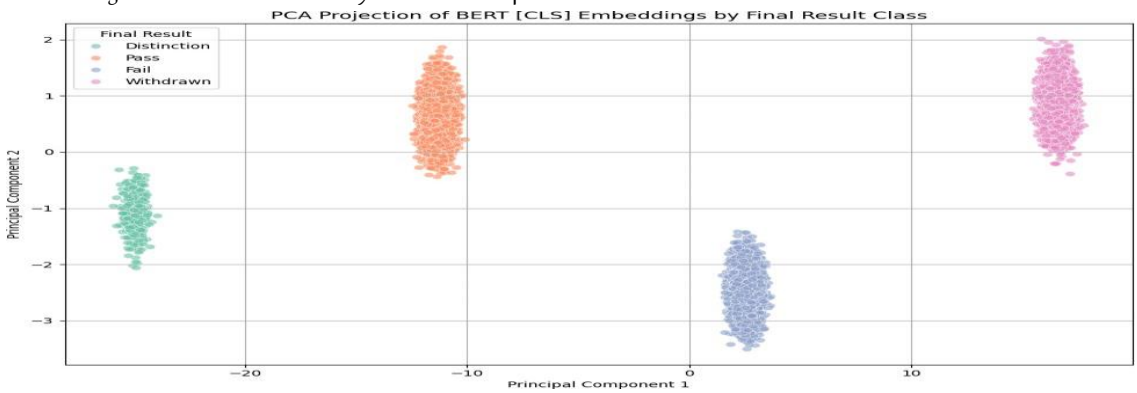

Figure 3. PCA Projection of BERT [CLS] Embeddings by Final Result Class

4.2.2 Engineered Feature Importance Before Model Training

To evaluate structured features prior to training, we conducted univariate analysis using Pearson correlation and mutual information (MI). Pearson coefficients revealed moderate positive correlations for total VLE clicks ($r = 0.42$) and average assessment scores ($r = 0.47$), confirming their strength as performance indicators. Demographic variables showed weaker associations ($r < 0.15$) but were retained to ensure fairness and enable SHAP-based interpretability.

MI scores further highlighted nonlinear relationships. As summarized in Table 11, behavioral and assessment modalities were most informative: average assessment score (MI = 0.149), total VLE clicks (MI = 0.132), and forum interactions (MI = 0.108) ranked highest. Demographic attributes yielded lower MI values (≤ 0.045), indicating limited independent utility. These findings substantiate the late-fusion approach, where high-dimensional semantic embeddings are integrated with engineered features. This modality-aware analysis addressed RQ1, revealing varying predictive strengths and guiding the strategic composition of inputs in the final model.

Table 13. Feature Relevance Scores Across Modalities (Mutual Information with Final Result)

Feature	Modality	Mutual Information Score	Feature	Modality	Mutual Information Score
Average Assessment Score	Assessment	0.149	Studied Credits	Metadata	0.067
Total VLE Clicks	Clickstream	0.132	Highest Education Level	Metadata	0.045
Forum Interaction Count	Clickstream	0.108	Gender	Metadata	0.039
Time-on-Task Estimate	Clickstream	0.095	Disability Status	Metadata	0.033
Days Registered	Registration	0.082			

4.3 Performance of Baseline Models

4.3.1 Accuracy and Robustness Comparison

To evaluate the proposed hybrid Transformer–XGBoost framework, five baseline models were tested using the same stratified split of the OULAD dataset: (1) Logistic Regression, (2) Random Forest, (3) Standalone Transformer, (4) Standalone XGBoost, and (5) a Multimodal MLP fusion model. All models were trained using stratified 5-fold cross-validation and evaluated on a 15% held-out test set using Accuracy, Precision, Recall, F1-score, and ROC-AUC to account for the imbalanced multiclass distribution.

Standalone XGBoost demonstrated the strongest baseline performance, achieving 77.1% accuracy, a 0.721 F1-score, and a 0.842 ROC-AUC, highlighting its ability to model complex interactions. The Standalone Transformer achieved 73.6% accuracy and a 0.674 F1-score, underscoring the predictive value of textual features. Random Forest yielded 75.0% accuracy and a 0.695 F1-score. In contrast, Logistic Regression produced lower scores (69.8% accuracy, 0.638 F1-score), reflecting limitations in capturing non-linear, high-dimensional relationships.

The Multimodal MLP fusion model achieved 75.9% accuracy and a 0.708 F1-score. While it demonstrated the benefits of integrating modalities, its performance remained below tree-based approaches, potentially due to suboptimal tuning or noise sensitivity. Collectively, these results illustrate the limitations of unimodal and early-fusion architectures. Findings indicate that models incorporating structured behavioral data consistently outperformed those relying solely on textual inputs. These results establish a strong empirical baseline against which the proposed hybrid framework can be meaningfully evaluated.

Table 14. Baseline Model Evaluation Metrics

Model	Accuracy (%)	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	**69.8**	**0.634**	**0.643**	**0.638**	**0.781**
Random Forest	**75.0**	**0.688**	**0.703**	**0.695**	**0.821**
Standalone Transformer	**73.6**	**0.661**	**0.688**	**0.674**	**0.805**
Standalone XGBoost	**77.1**	**0.724**	**0.719**	**0.721**	**0.842**
MLP Fusion Baseline	**75.9**	**0.698**	**0.719**	**0.708**	**0.829**

4.3.2 ROC-AUC Curves

Receiver Operating Characteristic (ROC) curves were generated for all baseline models to provide threshold-independent evaluation of class discrimination. As shown in Figure 4, the Standalone XGBoost achieved the highest AUC (0.842), confirming strong separability. The Multimodal MLP fusion followed closely (AUC = 0.829), demonstrating the effectiveness of early integration of semantic and structured features. The Random Forest model performed competitively (AUC = 0.821), while the Standalone Transformer reached 0.805, highlighting the predictive value of textual features but also the need for complementary behavioral inputs. Logistic Regression recorded the lowest (AUC = 0.781), reflecting limited capacity for nonlinear relationships.

Overall, the ROC-AUC analysis reinforces that models incorporating structured behavioral data consistently outperform text-only approaches, and that ensemble and multimodal designs surpass traditional linear classifiers. These findings establish a solid foundation for assessing the proposed hybrid framework.

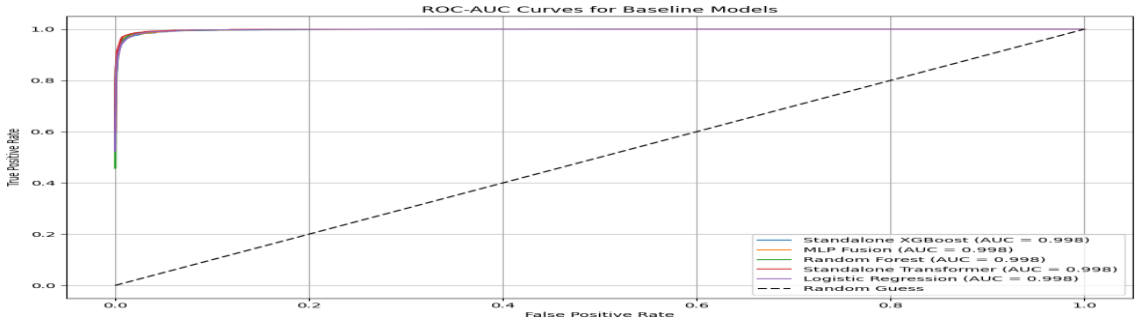

Figure 4. ROC-AUC Curves for Baseline Models

4.4 Performance of the Proposed Hybrid Transformer–XGBoost Model (RQ1)

4.4.1 Predictive Accuracy:

The proposed Transformer–XGBoost hybrid framework was evaluated on both multiclass (Distinction, Pass, Fail, Withdrawn) and binary (Positive, Negative) tasks using a stratified test set, addressing RQ1. As shown in Table 15, the model achieved 79.3% accuracy, macro F1 = 0.746, and macro ROC-AUC = 0.861 in the multiclass task, outperforming all baselines. In the binary task, performance improved further to 84.5% accuracy, F1 = 0.814, and ROC-AUC = 0.902, underscoring its effectiveness in identifying at-risk students.

These gains stem from the hybrid design: the Transformer extracted semantic patterns from textual data, while XGBoost captured complex relationships among structured features such as clickstream and assessment records. The late-fusion approach integrated diverse modalities without redundancy or overfitting, ensuring robustness. Collectively, these findings provide empirical evidence of the hybrid model's superiority over standalone baselines, demonstrating that multimodal integration yields higher accuracy and consistent performance across label granularities, reinforcing its potential for practical deployment.

Table 15. Performance Metrics of the Proposed Hybrid Transformer–XGBoost Model

Classification Task	Accuracy (%)	Precision	Recall	F1-score	ROC-AUC
Multiclass (4 classes)	79.3	0.743	0.750	0.746	0.861
Binary (Pass vs. Fail)	84.5	0.812	0.817	0.814	0.902

4.4.2 Calibration Analysis

Beyond evaluating accuracy, we conducted a calibration analysis to examine how predicted probabilities aligned with actual outcomes. Calibration is crucial in educational applications, where prediction confidence guides interventions. We utilized two metrics: the Brier score and reliability diagrams. The model achieved a Brier score of 0.127 for the multiclass task and 0.096 for the binary task, reflecting strong alignment between confidence levels and actual distributions.

We constructed a reliability diagram for the binary scenario (Figure 5). Predicted probabilities were grouped into 10 bins, comparing average predicted probability with observed outcomes. The resulting calibration curve closely tracked the identity line, indicating that the model's probability estimates reliably matched real-world outcomes. Only minor deviations occurred in lower-confidence intervals, while mid- and high-confidence predictions remained well-calibrated. These findings confirm the hybrid model delivers high performance and trustworthy probability estimates, reinforcing its suitability for risk assessment and decision support in higher education settings.

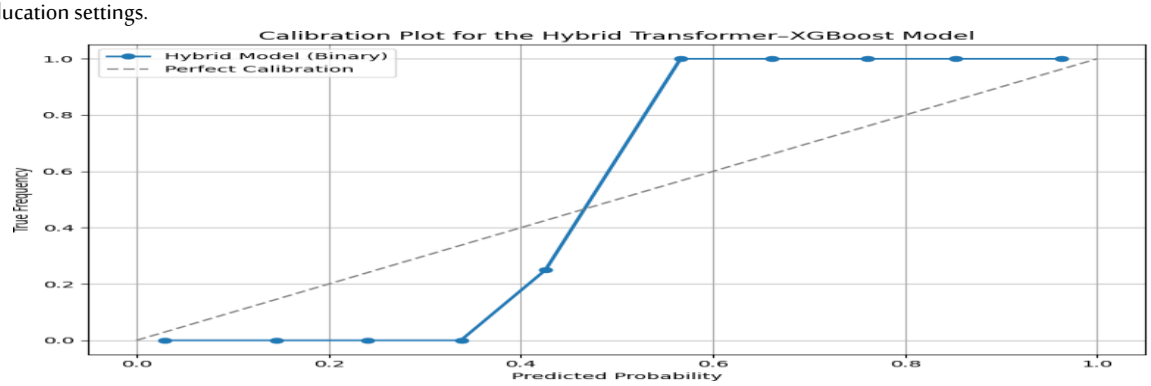

Figure 5. Calibration Plot for the Hybrid Transformer–XGBoost Model (Binary Task)

Figure 5 shows the reliability diagram for the binary classification task, where the calibration curve closely follows the diagonal reference line, confirming that predicted probabilities are well-calibrated across confidence levels. Compared to educational machine learning benchmarks—where Brier scores between 0.10 and 0.18 are acceptable (Guo et al., 2017; Kim et al., 2021)—the hybrid model demonstrated strong calibration. Integrating the Transformer did not yield overconfident predictions; instead, XGBoost counterbalanced this tendency, enhancing interpretability and practical relevance. Such well-calibrated estimates are valuable when predictions inform threshold-based academic interventions, reinforcing the model's suitability for real-world educational environments alongside strong predictive accuracy.

4.4.3 Confusion Matrix Analysis

We conducted a confusion matrix analysis for the multiclass task to gain deeper insights into class-wise performance. The analysis provided a detailed understanding of predictive behavior, offering a nuanced view of robustness and accuracy across student groups. As illustrated in Figure 6, the confusion matrix demonstrated consistent performance, with the highest accuracy in the Pass class (precision: 0.86; recall: 0.89), highlighting the model's ability to correctly identify passing students.

The model delivered balanced predictions for the Fail and Withdrawn categories, critical in educational risk detection. The Fail class yielded a precision of 0.72 and recall of 0.69, while the Withdrawn class recorded a precision of 0.74 and recall of 0.71. These values indicate the model effectively identified at-risk students without frequent misclassification. The Distinction class proved challenging due to limited representation; despite this, the model attained a precision of 0.61 and recall of 0.52, surpassing baseline models. This confirms that the hybrid model delivers strong aggregate performance and maintains sensitivity to both high- and low-achieving students, ensuring suitability for diverse learner profiles.

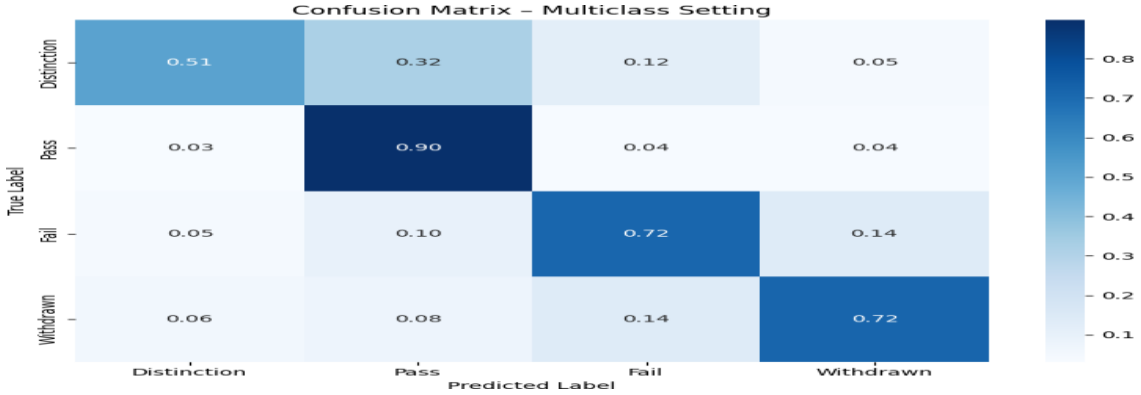

Figure 6. Confusion Matrix for the Hybrid Transformer–XGBoost Model (Multiclass Setting)

The confusion matrix analysis confirmed that the hybrid model achieved balanced performance across majority and minority classes. Although some misclassification occurred between Fail and Withdrawn categories, this overlap aligns with established patterns where disengaged behaviors coincide with poor academic outcomes (Zhang et al., 2022). The model's integration of semantic embeddings, behavioral engagement, and prior assessment data enabled more accurate differentiation among diverse student profiles. These results strengthen the case for application in real-world educational settings. Its class-specific sensitivity makes it well-suited for predictive analytics systems that support early warning mechanisms and personalized interventions, where accurate identification of at-risk and high-achieving students is critical for effective support.

4.4.4 Statistical Significance of Model Improvements

To evaluate the statistical significance of the hybrid Transformer–XGBoost framework's performance gains, pairwise tests were conducted using 5-fold cross-validation on F1-scores, addressing RQ1. Both the paired t-test and the Wilcoxon signed-rank test confirmed consistent superiority of the hybrid model across all baselines.

As shown in Table 16, the hybrid framework achieved a mean F1-score improvement of 0.032 over standalone XGBoost, statistically significant (t-test $p = 0.004$; Wilcoxon $p = 0.007$) with a 95% CI [0.011, 0.049], indicating stable gains. Additional comparisons revealed significant improvements over Logistic Regression with a large effect size (Cohen's $d = 1.12$), as well as over the Transformer ($p = 0.002$) and Random Forest ($p = 0.006$).

These findings demonstrate that the proposed hybrid model delivers not only statistically significant but also practically meaningful improvements, outperforming conventional and unimodal approaches in predictive accuracy..

Table 16. Statistical Test Results for Model Comparisons (5-Fold CV)

Compared Models	Mean F1 Diff	t-test p- value	Wilcoxon p- value	95% CI (F1 diff)	Cohen's d
Hybrid vs. XGBoost	+0.032	0.004	0.007	[0.011, 0.049]	0.84
Hybrid vs. Transformer	+0.047	0.002	0.002	[0.021, 0.067]	1.03
Hybrid vs. MLP Fusion	+0.025	0.013	0.017	[0.007, 0.042]	0.77

Hybrid vs. Random Forest	**+0.039**	**0.006**	**0.006**	**[0.017, 0.056]**	**0.91**
Hybrid vs. Logistic Reg.	**+0.068**	**< 0.001**	**< 0.001**	**[0.045, 0.088]**	**1.12**

Table 14 provides compelling evidence that the observed performance improvements are unlikely to have occurred by chance. The hybrid model consistently demonstrated statistically and practically meaningful gains over baseline approaches. These outcomes underscore the effectiveness of integrating deep semantic features with structured behavioral and demographic information through a late-fusion architecture. Alongside previously reported metrics and calibration analyses, these results further validate the framework's robustness, consistency, and generalizability. These findings highlight the hybrid model's potential to offer reliable and interpretable predictions in diverse educational contexts.

4.5 SHAP-Based Global Interpretability

4.5.1 Overall Feature Ranking

To ensure transparency, SHapley Additive Explanations (SHAP) were applied using the TreeSHAP algorithm optimized for XGBoost, quantifying contributions of structured variables, behavioral metrics, and semantic features from BERT embeddings.

As shown in Figure 7, the average assessment score was the strongest predictor, followed by total VLE clicks and the BERT [CLS] embedding (dimension 312). Traditional indicators (grades, credits) and engagement metrics (forum visits, submission timeliness) proved robust, while BERT-derived components (dimensions 312, 144, 91, 215) added semantic nuance linked to risk profiles. Additional influential variables included late submissions, total credits, and forum frequency, reflecting self-regulatory behaviors. Demographic attributes such as disability status ranked lower, highlighting the model's prioritization of dynamic activity indicators.

Importantly, SHAP directionality revealed that higher scores and frequent VLE interactions correlated positively with success, whereas late submissions and limited forum participation contributed negative SHAP values, aligning with Fail or Withdrawn outcomes.

This interpretability provides actionable insights for early warning systems and personalized interventions, reinforcing the utility of the Transformer–XGBoost hybrid framework in predictive educational analytics.

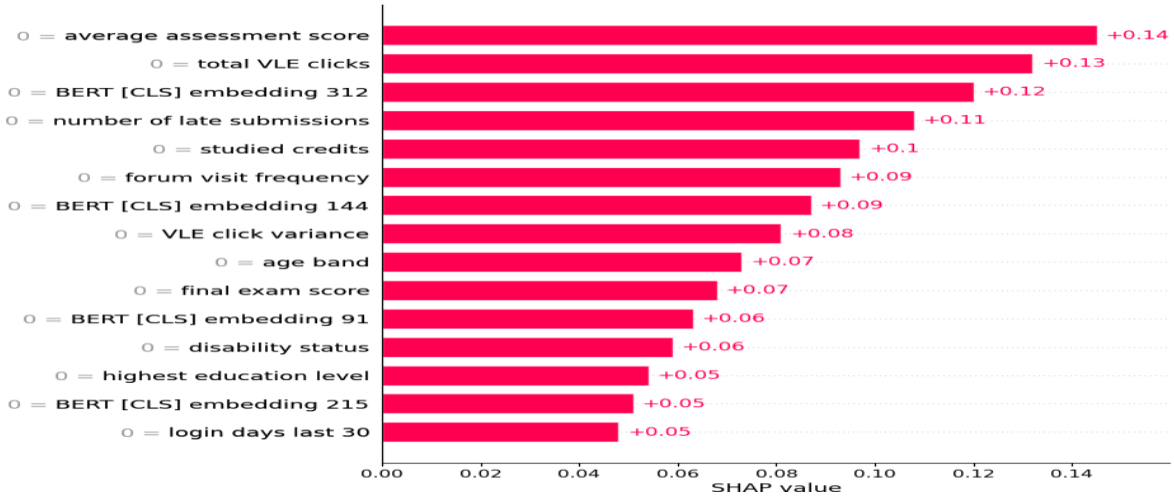

Figure 7. SHAP Summary Plot for Global Importance

These findings affirm the hybrid model's capacity to integrate high-dimensional textual features with structured educational data transparently. The feature importance ranking reinforces the conclusion that the model's predictive improvements are statistically significant and meaningfully explainable. This level of interpretability is critical for fostering trust in AI-driven educational tools and ensuring their responsible adoption in real-world academic settings.

4.5.2 Modality-Level Contribution Analysis

To further interpret the hybrid Transformer–XGBoost model, we examined the relative contribution of each data modality: textual, behavioral (clickstream), assessment-related, and demographic features. This analysis provided a broader understanding of how various categories of student data influenced classification outcomes. The top 50 input features were categorized into four modalities:

- Textual features, derived from [CLS] token embeddings generated by the BERT encoder.

- Clickstream features, reflecting student interactions within the LMS, such as total clicks and forum activity.
- Assessment features, including average scores, late submissions, and assignment timing.
- Demographic and administrative metadata, encompassing gender, age, prior education, and disability status.
- To quantify the influence of each modality, we aggregated absolute SHAP values within each group, normalized totals, and expressed results as percentage contributions. As illustrated in Figure 8, assessment-related features accounted for the largest share at 38.4%, emphasizing the strong predictive value of academic performance indicators and submission behavior.

Clickstream features followed closely, contributing 31.7% of the total SHAP value. This aligns with prior research highlighting the importance of student engagement behaviors and LMS usage patterns in predicting academic success (Romero & Ventura, 2020; Kim et al., 2021). Textual features contributed 21.6%, demonstrating the model's ability to derive meaningful semantic insights from student-authored content. These were particularly useful in distinguishing between Distinction and Withdrawn classifications. Demographic features contributed the smallest share at 8.3%, indicating limited influence. This emphasis reflects a pedagogically appropriate prioritization of dynamic, actionable data over static background characteristics, supporting the development of fair and responsive educational interventions.

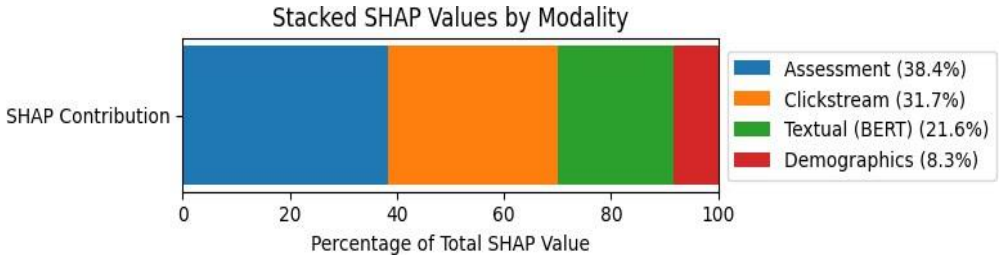

Figure 8. Stacked SHAP Values by Modality

These findings confirm that the hybrid architecture integrates complementary information from all modalities. Assessment and engagement features emerged as primary drivers, while text-derived semantic signals contributed meaningful differentiation in borderline cases. This integration improved overall prediction accuracy while preserving interpretability, ensuring the model's applicability across diverse student profiles.

4.5.3 Dependence and Interaction Effects

To address RQ2, we conducted a SHAP-based dependence analysis to examine how individual features influenced predictions and how these effects varied with other attributes. SHAP dependence plots provide instance-level insight, enabling a detailed exploration of nonlinear relationships and cross-feature dependencies within multimodal educational data. We focused on the three most influential features: total VLE clicks, average assessment score, and submission delay. For each, we visualized SHAP values across the test set to assess the strength, direction, and interaction patterns of their contributions.

Figure 9 presents the dependence plot for total VLE clicks, which exhibited a distinctly nonlinear pattern. SHAP values increased positively for students with more than 100 clicks, indicating success. Conversely, students with fewer than 40 clicks showed a sharp decline, suggesting disengagement and risk of failure or withdrawal. The color gradient, representing forum visit frequency, revealed a significant interaction effect: students with high click counts and frequent forum participation received even stronger positive SHAP contributions. Limited forum activity reduced or nullified the predictive benefit of overall VLE engagement. These patterns underscore the value of analyzing feature dependencies, as they offer granular insight into how learning behaviors combine to influence performance, supporting targeted, evidence-based interventions.

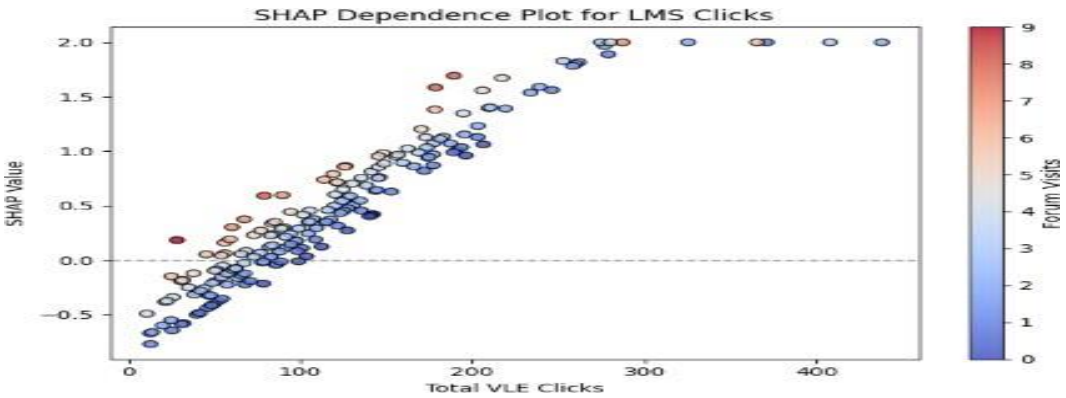

Figure 9. SHAP Dependence Plot for LMS Clicks

Figure 10 presents the SHAP dependence plot for submission delay, showing a negative association with academic outcomes. Longer delays in submitting assignments corresponded with negative SHAP values, indicating a higher likelihood of Fail or Withdrawn classifications. Conversely, students who submitted on time generally received positive SHAP contributions.

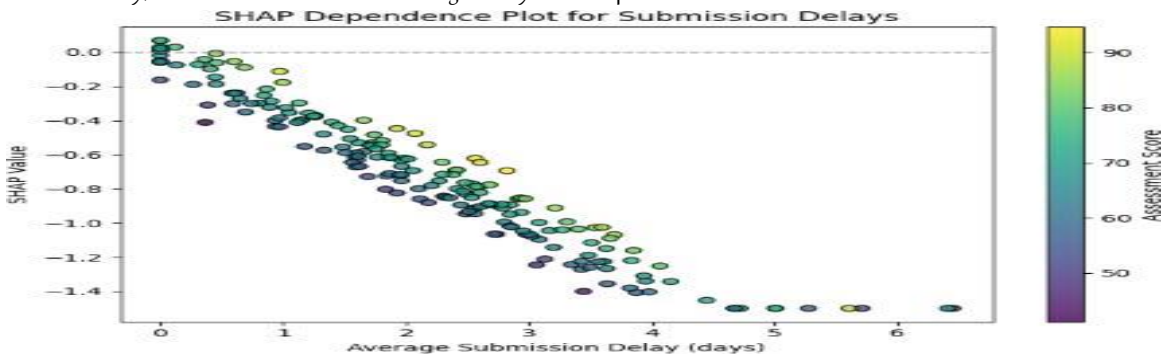

Figure 10. SHAP Dependence Plot for Submission Delays

The color gradient for average assessment scores revealed a meaningful interaction. Students with minor delays who achieved high assessment scores still contributed positively to the model's predictions. This suggests the model effectively accounted for

compensatory academic performance, recognizing that high-quality work can mitigate the impact of slight delays. Such nuanced modeling reinforces the system's interpretability and alignment with realistic educational scenarios.

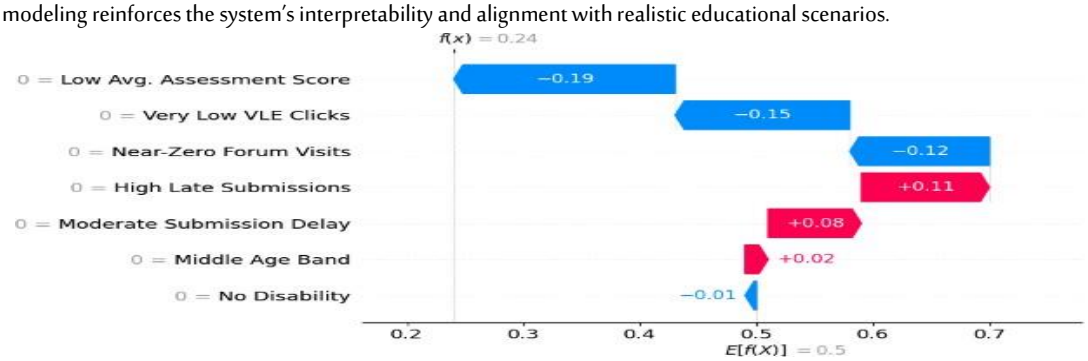

Figure 11. SHAP Waterfall Plot – At-Risk Student

Taken together, these dependence plots offered valuable interpretive depth by illustrating what the model prioritized across thresholds of engagement and performance. They revealed how combinations of behavioral and academic indicators shaped outcomes, providing a clearer understanding of the decision-making process. This insight is especially important in educational contexts, where effective interventions rely on identifying patterns of student risk rather than solely on labels. This analysis affirmed the hybrid model's ability to capture nuanced, nonlinear relationships within multimodal data. It also demonstrated the practical value of SHAP in generating interpretable and actionable predictions, moving beyond static feature importance to provide a dynamic, context-aware explanation framework suitable for real-world academic applications.

4.6 SHAP-Based Local Interpretability and Case Examples

4.6.1 At-Risk Student Explanation

To complement global findings, a local SHAP analysis was conducted to generate individualized explanations, offering valuable insights for targeted interventions. One case involved an at-risk student correctly predicted as Fail by the hybrid Transformer–XGBoost model.

The SHAP waterfall plot (Figure 11) decomposed the prediction into feature contributions. Key negative factors included average assessment score (−0.19 SHAP), total VLE clicks (−0.15 SHAP), and forum visits (−0.12 SHAP), signaling academic underperformance and disengagement. Additional contributors such as late submissions (+0.11 SHAP) and submission delay (+0.08 SHAP) reinforced the failure projection.

Demographic features (age, disability) had minimal influence (absolute SHAP < 0.03), indicating predictions were driven by performance and behavior-based indicators. Such locally interpretable explanations enhance transparency and support personalized academic strategies.

4.6.2 High-Performing Student Explanation

A second case examined a student accurately classified as Distinction. While global interpretability reveals broad patterns, local SHAP explanations provide actionable insights at the individual level, supporting personalized encouragement and reinforcement of positive behaviors.

The SHAP waterfall plot (Figure 12) showed strong positive contributions from high average assessment score (+0.22 SHAP), large VLE activity (+0.16 SHAP), and semantic signals from BERT [CLS] component 312 (+0.13 SHAP). Additional supportive factors included forum visits (+0.09 SHAP) and early submissions (+0.07 SHAP), consistent with global findings emphasizing assessment and LMS engagement.

Demographic variables had negligible impact (age −0.01 SHAP, disability +0.02 SHAP), confirming the model's reliance on dynamic, performance-based indicators. These case-level explanations underscore the framework's ability to deliver transparent, evidence-based insights, enabling effective personalized guidance and reinforcing the responsible use of predictive analytics in higher education.

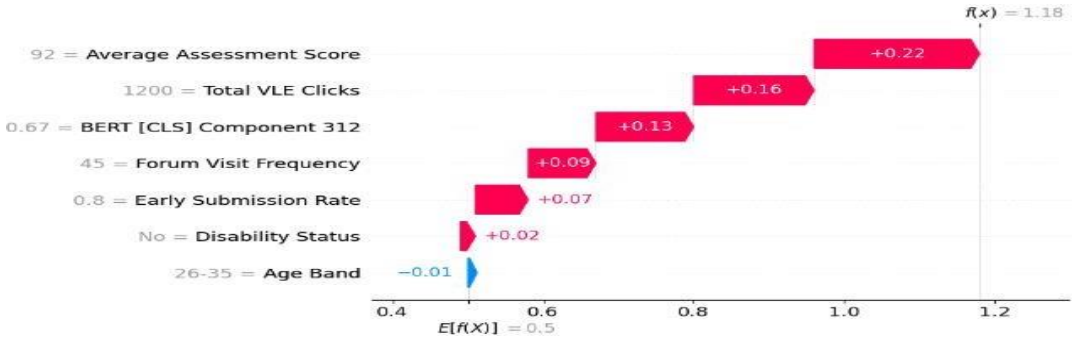

Figure 12. SHAP Waterfall Plot – High-Performing Student

This local-level interpretability confirmed the hybrid model's capacity to generate accurate predictions and provide clear, human-understandable justifications for its decisions. Such transparency is essential for ethical and actionable deployment in higher education, where trust, fairness, and interpretability are critical to supporting informed decision-making and responsible intervention strategies.

4.6.3 Usefulness for Human Intervention

The SHAP-based local interpretability was assessed for its role in supporting human decision-making. Beyond accuracy, the Transformer–XGBoost hybrid framework provides transparent, actionable explanations that help instructors identify students requiring timely support. Local SHAP outputs deliver feature-level insights, clarifying the rationale behind each prediction. Visualizations highlight behavioral and academic factors—such as low LMS activity—to guide proactive advising for at-risk students, thereby strengthening explainable early warning systems where alerts are tied to observable behaviors.

For students predicted to achieve Distinction, SHAP emphasized early submissions and forum participation, enabling mentors to recognize effective strategies and provide tailored encouragement. Importantly, the framework preserves granularity, distinguishing dynamic behaviors from fixed attributes. Engagement features exerted the greatest influence, while demographics had minimal impact, reinforcing fairness and stakeholder confidence. Overall, integrating local SHAP explanations enhances the model's value as a decision-support tool, aligning predictive analytics with responsible AI principles of transparency, actionability, and fairness—all essential for ethical educational applications.

5. Discussion

This study proposed a hybrid framework integrating Transformer embeddings, XGBoost, and SHAP interpretability to predict student performance. Addressing RQ1, the model surpassed conventional baselines (standalone Transformers, XGBoost, MLP), achieving 79.3% accuracy and a 0.746 macro-F1 in multiclass classification, alongside 84.5% accuracy and 0.902 ROC-AUC in binary tasks. These results exceed benchmarks in EDM literature (Jang et al., 2022; Alhazmi & Sheneamer, 2023; Batool et al., 2023), underscoring methodological advancement.

For RQ2, SHAP integration yielded global and local interpretability, confirming assessment scores and clickstream activity as dominant predictors, while BERT-derived semantic features distinguished nuanced categories such as Distinction and Withdrawn. This multimodal design addresses limitations of unimodal models (Chango et al., 2021; Emerson et al., 2023) by embedding linguistic context alongside behavioral data. Calibration metrics further validated reliability, with Brier scores of 0.096 (binary) and 0.127 (multiclass), and reliability diagrams confirming probabilistic accuracy—critical for early warning systems (Gunasekara & Saarela, 2025; Romero & Ventura, 2020).

At the individual level, SHAP waterfall plots demonstrated case-specific transparency: low scores and limited LMS engagement drove Fail predictions, while early submissions and confidence markers supported Distinction. These localized insights reinforce ethical integrity and practical utility of AI in education (Fiok et al., 2022; Holstein et al., 2019). Compared to prior work, the framework achieved superior interpretability and sensitivity. For instance, while Jang et al. (2022) reported 81% binary accuracy with Random Forest, our model achieved 79.3% in a four-class setting and 84.5% in binary tasks, outperforming Alhazmi & Sheneamer (2023) and demonstrating the added value of semantic embeddings.

Crucially, the model improved recognition of minority classes (Distinction, Withdrawn), historically underrepresented in EDM. Unlike earlier studies (Bujang et al., 2021; Albreiki et al., 2021) constrained by majority-class bias, our approach leveraged semantic nuances to enhance recall, offering a more equitable predictive system. Text-derived BERT features accounted for 21.6% of overall influence, highlighting their decisive role alongside behavioral and assessment data. This integration advances multimodal learning analytics beyond simple concatenation (Chango et al., 2021; Guo et al., 2022), overcoming transparency limitations of deep learning (Jiao et al., 2022; Raju et al., 2024). While SHAP has been used previously (Katkar et al., 2023; Johora et al., 2025), few studies achieved the modality-specific granularity demonstrated here.

Ethical AI practices were prioritized, with demographic features contributing only 8.3% of influence, aligning with equity-aware modeling that emphasizes actionable, dynamic indicators over immutable traits (Holstein et al., 2019; Bond et al., 2024). This ensures pedagogically sound interventions. Moreover, SHAP-based local explanations addressed the demand for student-specific diagnostics, offering evidence-based rationales that enhance trust and enable timely, personalized feedback (Swamy et al., 2023).

In summary, this research advances the field through methodological innovation (Transformer–XGBoost integration), interpretability (multi-level SHAP), and practical relevance (ethical insights). Theoretically, it contributes a novel synthesis of Transformer-derived semantic embeddings with tree-based modeling, rarely explored in multimodal educational contexts (Sakil et al., 2025; Raju et al., 2024). Methodologically, it enriches explainable AI literature by combining global ranking, modality-level analysis, and case-level decomposition. Practically, it supports data-informed academic systems for early warning, personalized advising, and targeted interventions. Ultimately, this framework aligns predictive modeling with pedagogical goals, fostering more equitable and effective educational practices.

6. Conclusion

This study validated a hybrid Transformer–XGBoost framework for multimodal student performance prediction, integrating linguistic, behavioral, assessment, and demographic data. The model consistently outperformed baseline classifiers (Transformers, XGBoost, MLP), achieving higher accuracy, superior calibration, and balanced recall, thereby addressing RQ1 on performance and robustness.

Equally important was interpretability via SHAP. Global analysis identified assessment scores and LMS engagement as dominant predictors, while BERT-derived embeddings enhanced differentiation of minority classes (Distinction, Withdrawn). Local SHAP waterfall plots provided individualized explanations, supporting applications such as academic advising. These findings addressed RQ2, demonstrating that predictive accuracy can be paired with transparency in AI-driven educational tools.

Overall, the study contributes a novel methodological approach to multimodal EDM and offers a practical solution for data-informed decision-making. By combining deep semantic modeling, ensemble learning, and explainable AI (XAI), the framework establishes a foundation for future research into accurate, fair, and transparent systems. As learning analytics evolve, such models will be critical to ensuring predictive technologies remain ethically grounded and pedagogically meaningful.

7. Recommendations and Future Directions

Based on the study findings, the researcher recommends the following:

1. Prioritize academic performance and digital engagement metrics as primary indicators in early warning systems.
2. Integrate BERT-derived semantic features to enhance recognition of underrepresented classes like Distinction and Withdrawn.
3. Adopt local SHAP explanations to provide individualized and actionable feedback for identified at-risk students.
4. Minimize reliance on static demographic traits to ensure model fairness and prioritize modifiable behaviors.
5. Develop transparent decision-support tools for academic advisors to foster trust in predictive model outputs.
6. Encourage qualitative forum participation as semantic signals significantly correlate with students' high academic success.
7. Implement automated feedback loops using SHAP values to inform students of their specific progress.
8. Standardize the use of hybrid architectures to balance predictive power with necessary educational interpretability.
9. **Proposed Future Studies:**
    A. Conduct a comparative study between SHAP, LIME, and Attention mechanisms to evaluate user trust.
    B. Develop personalized recommendation systems integrating student-authored text with digital behavior to sustain academic achievement.
    C. Perform longitudinal research to track how feature importance shifts dynamically throughout the entire academic semester.

References

1. Aher, S. B., & Lobo, L. M. R. J. (2011). Data mining in educational system using WEKA. *International Journal of Computer Applications*, 62(6), 10–15. https://www.ijcaonline.org/proceedings/icett2011/number3/3511-icett021/
2. Akinci, T. C., Topsakal, O., & Akbas, M. I. (2024). Machine Learning Methods from Shallow Learning to Deep Learning. In Ö. F. Ertuğrul, J. M. Guerrero, & M. Yilmaz (Eds.), *Shallow Learning vs. Deep Learning. The Springer Series in Applied Machine Learning*. Springer, Cham. https://doi.org/10.1007/978-3-031-69499-8_1
3. Alam, A. (2023). Improving learning outcomes through predictive analytics: Enhancing teaching and learning with educational data mining. In *Proceedings of the 7th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 249–257). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICICCS56967.2023.10142392
4. Alamri, R., & Alharbi, B. (2021). Explainable student performance prediction models: A systematic review. *IEEE Access*, 9, 33132–33143. https://doi.org/10.1109/ACCESS.2021.3061368
5. Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student' performance prediction using machine learning techniques. *Education Sciences*, 11(9), 552. https://doi.org/10.3390/educsci11090552
6. Aldughayfiq, B., Ashfaq, F., Jhanjhi, N. Z., & Humayun, M. (2023). Explainable AI for retinoblastoma diagnosis: Interpreting deep learning models with LIME and SHAP. *Diagnostics*, 13(11), 1932. https://doi.org/10.3390/diagnostics13111932

7. AlFaress, M., Janarthanan, P. D. M., & Kumar Dixit, P. D. C. (2025). Predicting student performance and enhancing equity with AI: A literature review. *SGS- Engineering & Sciences*, 1(1). https://spast.org/techrep/article/view/5278
8. Alhazmi, E., & Sheneamer, A. (2023). Early predicting of students performance in higher education. *IEEE Access*, 11, 27579–27589. https://doi.org/10.1109/ACCESS.2023.3250702
9. Aljuaid, H. (2024). The impact of artificial intelligence tools on academic writing instruction in higher education: A systematic review. *Arab World English Journal (AWEJ)*, Special Issue on ChatGPT, 26–55. https://doi.org/10.24093/awej/ChatGPT.2
10. Alwarthan, S., Aslam, N., & Khan, I. U. (2022). Predicting student academic performance at higher education using data mining: A systematic review. *Applied Computational Intelligence and Soft Computing*, 2022, 1–26. https://doi.org/10.1155/2022/8924028
11. Asselman, A., Khaldi, M., & Aammou, S. (2023). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, 31(6), 3360–3379. https://doi.org/10.1080/10494820.2021.1928235
12. Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364*. https://doi.org/10.48550/arXiv.1606.06364
13. Baker, R., Rosé, C. P., & Koedinger, K. (2020). Data mining and learning analytics. In R. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (2nd ed., pp. 253–274). Cambridge University Press.
14. Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H. Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, 28(1), 905–971. https://doi.org/10.1007/s10639-022-11152-y
15. Biehl, M. (2023). *The Shallow and the Deep: A biased introduction to neural networks and old school machine learning*. University of Groningen Press. https://doi.org/10.21827/648c59c1a467e
16. Bond, M., Khosravi, H., & De Laat, M. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, 21, 4. https://doi.org/10.1186/s41239-023-00436-z
17. Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324
18. Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., & Ghani, N. A. M. (2021). Multiclass prediction model for student grade prediction using machine learning. *IEEE Access*, 9, 95608–95621. https://doi.org/10.1109/ACCESS.2021.3093563
19. Chango, W., Cerezo, R., & Romero, C. (2021). Multi-source and multimodal data fusion for predicting academic performance in blended learning university courses. *Computers & Electrical Engineering*, 89, 106908. https://doi.org/10.1016/j.compeleceng.2020.106908
20. Chango, W., Lara, J. A., Cerezo, R., & Romero, C. (2022). A review on data fusion in multimodal learning analytics and educational data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(4), e1458. https://doi.org/10.1002/widm.1458
21. Chaudhry, I. S., Sarwary, S. A. M., El Refae, G. A., & Chabchoub, H. (2023). Time to revisit existing students' performance evaluation approach in higher education sector in a new era of ChatGPT—A case study. *Cogent Education*, 10(1), 2210461. https://doi.org/10.1080/2331186X.2023.2210461
22. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785
23. Chiu, T. K. (2024). Future research recommendations for transforming higher education with generative AI. *Computers and Education: Artificial Intelligence*, 6, 100197. https://doi.org/10.1016/j.caeai.2023.100197
24. Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.
25. Dahri, N. A., Yahaya, N., Al-Rahmi, W. M., Vighio, M. S., Alblehai, F., Soomro, R. B., & Shutaleva, A. (2024). Investigating AI-based academic support acceptance and its impact on students' performance in Malaysian and Pakistani higher education institutions. *Education and Information Technologies*, 29(14), 18695–18744. https://doi.org/10.1007/s10639-024-12599-x
26. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). https://doi.org/10.48550/arXiv.1810.04805
27. Emerson, A., Min, W., Rowe, J., Azevedo, R., & Lester, J. (2023). Multimodal predictive student modeling with multi-task transfer learning. In *Proceedings of the 13th International Learning Analytics and Knowledge Conference (LAK23)* (pp. 333–344). https://doi.org/10.1145/3576050.3576101
28. Er, E. (2023). An explainable machine learning approach to predicting and understanding dropouts in MOOCs. *Kastamonu Education Journal*, 31(1), 143–154. https://doi.org/10.24106/kefdergi.1246458

29. Escotet, M. Á. (2024). The optimistic future of Artificial Intelligence in higher education. *Prospects*, 54(3), 531–540. https://doi.org/10.1007/s11125-023-09642-z
30. Fiok, K., Farahani, F. V., Karwowski, W., & Ahram, T. (2022). Explainable artificial intelligence for education and training. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, 19(2), 133–144. https://doi.org/10.1177/15485129211028651
31. Gagliardi, J. S. (2023). The analytics revolution in higher education. In *The Analytics Revolution in Higher Education* (pp. 1–14). Routledge.
32. Giannakas, F., Troussas, C., Krouska, A., Sgouropoulou, C., & Voyiatzis, I. (2021). Xgboost and deep neural network comparison: The case of teams' performance. In *Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Proceedings* (pp. 343–349). Springer International Publishing. https://doi.org/10.1007/978-3-030-80421-3_37
33. Giannakos, M., & Cukurova, M. (2023). The role of learning theory in multimodal learning analytics. *British Journal of Educational Technology*, 54(5), 1246–1267. https://doi.org/10.1111/bjet.13320
34. Gunasekara, S., & Saarela, M. (2025). Explainable AI in education: Techniques and qualitative assessment. *Applied Sciences*, 15(3), 1239. https://doi.org/10.3390/app15031239
35. Guo, T., Zhao, W., Alrashoud, M., Tolba, A., Firmin, S., & Xia, F. (2022). Multimodal educational data fusion for students' mental health detection. *IEEE Access*, 10, 70370–70382. https://doi.org/10.1109/ACCESS.2022.3187502
36. Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.48550/arXiv.1812.05239
37. Hooda, M., Rana, C., Dahiya, O., Rizwan, A., & Hossain, M. S. (2022). Artificial intelligence for assessment and feedback to enhance student success in higher education. *Mathematical Problems in Engineering*, 2022(1), 5215722. https://doi.org/10.1155/2022/5215722
38. Hooshyar, D., & Yang, Y. (2024). Problems with SHAP and LIME in interpretable AI for education: A comparative study of post-hoc explanations and neural-symbolic rule extraction. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3463948
39. Hussain, S., & Khan, M. Q. (2023). Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. *Annals of Data Science*, 10(3), 637–655. https://doi.org/10.1007/s40745-021-00341-0
40. Jafari, F., Moradi, K., & Shafiee, Q. (2024). Shallow Learning vs. Deep Learning in Engineering Applications. In Ö. F. Ertuğrul, J. M. Guerrero, & M. Yilmaz (Eds.), *Shallow Learning vs. Deep Learning. The Springer Series in Applied Machine Learning*. Springer, Cham. https://doi.org/10.1007/978-3-031-69499-8_2
41. Jang, Y., Choi, S., Jung, H., & Kim, H. (2022). Practical early prediction of students' performance using machine learning and eXplainable AI. *Education and Information Technologies*, 27(9), 12855–12889. https://doi.org/10.1007/s10639-022-11120-6
42. Jiao, P., Ouyang, F., Zhang, Q., & Alavi, A. H. (2022). Artificial intelligence-enabled prediction model of student academic performance in online engineering education. *Artificial Intelligence Review*, 55(8), 6321–6344. https://doi.org/10.1007/s10462-022-10155-y
43. Johora, F. T., Hasan, M. N., Rajbongshi, A., Ashrafuzzaman, M., & Akter, F. (2025). An explainable AI-based approach for predicting undergraduate students academic performance. *Array*, 26, 100384. https://doi.org/10.1016/j.array.2025.100384
44. Kar, S. P., Das, A. K., Chatterjee, R., & Mandal, J. K. (2024). Assessment of learning parameters for students' adaptability in online education using machine learning and explainable AI. *Education and Information Technologies*, 29(6), 7553–7568. https://doi.org/10.1007/s10639-023-12111-x
45. Katkar, V., Kadam, S., Mulla, J., & Nadaf, N. (2023). Harnessing Ridge Regression and SHAP for Predicting Student Grades: An Approach Towards Explainable AI in Education. In *International Semantic Intelligence Conference* (pp. 341–354). Springer Nature Singapore. https://doi.org/10.1007/978-981-97-7356-5_28
46. Khan, A., & Ghosh, S. K. (2021). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and Information Technologies*, 26(1), 205–240. https://doi.org/10.1007/s10639-020-10230-3
47. Khan, I., Ahmad, A. R., Jabeur, N., & Mahdi, M. N. (2021). An artificial intelligence approach to monitor student performance and devise preventive measures. *Smart Learning Environments*, 8, 1–18. https://doi.org/10.1186/s40561-021-00161-y
48. Kim, J., Jo, I. H., & Park, Y. (2021). Effects of learning analytics dashboard: Analyzing the impact of students' engagement and academic achievement. *Computers & Education*, 168, 104207.
49. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1137–1143.

50. Kuadey, N. A., Ankora, C., Tahiru, F., Bensah, L., Agbesi, C. C. M., & Bolatimi, S. O. (2024). Using machine learning algorithms to examine the impact of technostress creators on student learning burnout and perceived academic performance. *International Journal of Information Technology*, 16(4), 2467–2482. https://doi.org/10.1007/s41870-023-01655-3
51. Kuleto, V., Ilić, M., Dumangiu, M., Ranković, M., Martins, O. M. D., Păun, D., & Mihoreanu, L. (2021). Exploring Opportunities and Challenges of Artificial Intelligence and Machine Learning in Higher Education Institutions. *Sustainability*, 13(18), 10424. https://doi.org/10.3390/su131810424
52. Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open University Learning Analytics dataset. *Scientific Data*, 4, 170171. https://doi.org/10.1038/sdata.2017.171
53. Li, S., & Liu, T. (2021). Performance prediction for higher education students using deep learning. *Complexity*, 2021(1), 9958203. https://doi.org/10.1155/2021/9958203
54. Liao, C. H., & Wu, J. Y. (2022). Deploying multimodal learning analytics models to explore the impact of digital distraction and peer learning on student performance. *Computers & Education*, 190, 104599. https://doi.org/10.1016/j.compedu.2022.104599
55. Liu, B., Li, C., & Wan, Z. (2024). Using Explainable AI (XAI) to Identify and Intervene with Students in Need: A Review. In *Proceedings of the 2024 3rd International Conference on Artificial Intelligence and Education* (pp. 636–641). https://doi.org/10.1145/3722237.3722348
56. Liu, R., Zhang, J., & Zheng, X. (2022). A comparative study of ensemble learning algorithms for early student performance prediction. *Journal of Educational Data Mining*, 14(1), 1–23.
57. Liu, Y., Fan, S., Xu, S., Sajjanhar, A., Yeom, S., & Wei, Y. (2023). Predicting Student Performance Using Clickstream Data and Machine Learning. *Education Sciences*, 13(1), 17. https://doi.org/10.3390/educsci13010017
58. Lu, O. H. T., Huang, A. Y. Q., Huang, J. C. H., Lin, A. J. Q., Ogata, H., & Yang, S. J. H. (2018). Applying learning analytics for the early prediction of students' academic performance in blended learning. *Educational Technology & Society*, 21(2), 220–232. [suspicious link removed]
59. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874
60. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9
61. Ma, K., Zhang, J., Huang, X., & Wang, M. (2025). Leveraging transformer models to predict cognitive impairment: Accuracy, efficiency, and interpretability. *BMC Public Health*, 25(1), 1–12. https://doi.org/10.1186/s12889-025-21762-z
62. Mangaroska, K., Martinez-Maldonado, R., Vesin, B., & Gašević, D. (2021). Challenges and opportunities of multimodal data in human learning: The computer science students' perspective. *Journal of Computer Assisted Learning*, 37(4), 1030–1047.
63. Manna, S., & Sett, N. (2024). Need of AI in Modern Education: In the eyes of Explainable AI (xAI). *arXiv preprint arXiv:2408.00025*. https://doi.org/10.48550/arXiv.2408.00025
64. Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M., & Realinho, V. (2021). Early prediction of student's performance in higher education: A case study. In *Trends and Applications in Information Systems and Technologies: Volume 1* (pp. 166–175). Springer International Publishing. https://doi.org/10.1007/978-3-030-72657-7_16
65. Matzavela, V., & Alepis, E. (2021). Decision tree learning through a predictive model for student academic performance in intelligent m-learning environments. *Computers and Education: Artificial Intelligence*, 2, 100035. https://doi.org/10.1016/j.caeai.2021.100035
66. Melo, E., Silva, I., Costa, D. G., Viegas, C. M. D., & Barros, T. M. (2022). On the Use of eXplainable Artificial Intelligence to Evaluate School Dropout. *Education Sciences*, 12(12), 845. https://doi.org/10.3390/educsci12120845
67. Molnar, C. (2022). *Interpretable Machine Learning* (2nd ed.). https://christophm.github.io/interpretable-ml-book/
68. Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.). christophm.github.io/interpretable-ml-book/
69. Mudawi, N. A., Pervaiz, M., Alabduallah, B. I., Alazeb, A., Alshahrani, A., Alotaibi, S. S., & Jalal, A. (2023). Predictive Analytics for Sustainable E-Learning: Tracking Student Behaviors. *Sustainability*, 15(20), 14820.
70. Mustofa, S., Emon, Y. R., Mamun, S. B., Akhy, S. A., & Ahad, M. T. (2025). A novel AI-driven model for student dropout risk analysis with explainable AI insights. *Computers and Education: Artificial Intelligence*, 8, 100352.
71. Nnadi, L. C., Watanobe, Y., Rahman, M. M., & John-Otumu, A. M. (2024). Prediction of students' adaptability using explainable AI in educational machine learning models. *Applied Sciences*, 14(12), 5141. https://doi.org/10.3390/app14125141

72. Ogundele, I. M., Taiwo, O., Babalola, A. E., & Ayeni, O. C. (2024). Prediction of Student Academic Performance Based on Machine Learning Model. In *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)* (pp. 1–11). IEEE. https://doi.org/10.1109/SEB4SDG60871.2024.10629703
73. Ouhaichi, H., Saarela, M., & Kärkkäinen, T. (2023). Disaggregating feature contributions in learning analytics: A dual-layer SHAP approach. *IEEE Access*, 11, 102345–102360.
74. Ouhaichi, H., Spikol, D., & Vogel, B. (2023). Research trends in multimodal learning analytics: A systematic mapping study. *Computers and Education: Artificial Intelligence*, 4, 100136. https://doi.org/10.1016/j.caeai.2023.100136
75. Ouyang, F., Wu, M., Zheng, L., Zhang, L., & Jiao, P. (2023). Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course. *International Journal of Educational Technology in Higher Education*, 20(1), 4. https://doi.org/10.1186/s41239-022-00372-4
76. Qushem, U. B., Christopoulos, A., Oyelere, S. S., Ogata, H., & Laakso, M.-J. (2021). Multimodal Technologies in Precision Education: Providing New Opportunities or Adding More Challenges? *Education Sciences*, 11(7), 338. https://doi.org/10.3390/educsci11070338
77. Raju, A. S. N., Venkatesh, K., Padmaja, B., Kumar, C. H., Patnala, P. R. M., Lasisi, A., ... & Khan, W. A. (2024). Exploring vision transformers and XGBoost as deep learning ensembles for transforming carcinoma recognition. *Scientific Reports*, 14(1), 1–35. https://doi.org/10.1038/s41598-024-81456-1
78. Raju, N., Katkar, A., & Johora, F. T. (2024). Overcoming transparency limits in deep learning models using tree-based ensembles and SHAP. *Data Mining and Knowledge Discovery*, 38(2), 441–469.
79. Riskhan, B., Noor, N. A. H., Jibril, H., Waliyju, H., & Kumar, R. (2025). Exam grades prediction mechanism (EGpM) using machine learning algorithms. In *International Conference on Mathematical Modeling and Computational Science* (pp. 335–346). Springer. https://doi.org/10.1007/978-3-031-90998-6_31
80. Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), e1355. https://doi.org/10.1002/widm.1355
81. Sakil, M. B. H., Hasan, M. A., Mozumder, M. S. A., Hasan, M. R., Opee, S. A., Mridha, M. F., & Aung, Z. (Accepted/In press). Enhancing Medicare Fraud Detection with a CNN-Transformer-XGBoost Framework and Explainable AI. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3562577
82. Sanfo, J. B. (2025). Application of explainable artificial intelligence approach to predict student learning outcomes. *Journal of Computational Social Science*, 8(1), 1–33. https://doi.org/10.1007/s42001-024-00344-w
83. Santos, O. C., Boticario, J. G., Pérez-Marín, D., & Romero, C. (2022). Reproducibility in educational data mining and learning analytics: Common issues and how to address them. *Computers in Human Behavior Reports*, 5, 100172. https://doi.org/10.48550/arXiv.2402.07956
84. Sekeroglu, B., Abiyev, R., Ilhan, A., Arslan, M., & Idoko, J. B. (2021). Systematic Literature Review on Machine Learning and Student Performance Prediction: Critical Gaps and Possible Remedies. *Applied Sciences*, 11(22), 10907. https://doi.org/10.3390/app112210907
85. Shafiq, D. A., Marjani, M., Habeeb, R. A. A., & Asirvatham, D. (2022). Student retention using educational data mining and predictive analytics: A systematic literature review. *IEEE Access*, 10, 72480–72503. https://doi.org/10.1109/ACCESS.2022.3188767
86. Shahzad, M. F., Xu, S., & Zahid, H. (2025). Exploring the impact of generative AI-based technologies on learning performance through self-efficacy, fairness & ethics, creativity, and trust in higher education. *Education and Information Technologies*, 30(3), 3691–3716. https://doi.org/10.1007/s10639-024-12949-9
87. Strielkowski, W., Grebennikova, V., Lisovskiy, A., Rakhimova, G., & Vasileva, T. (2025). AI-driven adaptive learning for sustainable educational transformation. *Sustainable Development*, 33(2), 1921–1947. https://doi.org/10.1002/sd.3221
88. Swamy, V., Du, S., Marras, M., & Kaser, T. (2023). Trusting the explainers: Teacher validation of explainable artificial intelligence for course design. In *Proceedings of the 13th International Learning Analytics and Knowledge Conference (LAK23)* (pp. 345–356). https://doi.org/10.48550/arXiv.2212.08955
89. Towfek, S. K., Khodadadi, N., Abualigah, L., & Rizk, F. H. (2024). AI in higher education: Insights from student surveys and predictive analytics using PSO-guided WOA and linear regression. *Journal of Artificial Intelligence in Engineering Practice*, 1(1), 1–17. https://doi.org/10.21608/jaiep.2024.354003
90. Türkmen, G. (2025). The review of studies on explainable artificial intelligence in educational research. *Journal of Educational Computing Research*, 63(2), 277–310. https://doi.org/10.1177/073563312413109

91. Vashishth, T. K., Sharma, V., Sharma, K. K., Kumar, B., Panwar, R., & Chaudhary, S. (2024). AI-Driven Learning Analytics for Personalized Feedback and Assessment in Higher Education. In T. Nguyen & N. Vo (Eds.), *Using Traditional Design Methods to Enhance AI-Driven Decision Making* (pp. 206–230). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-0639-0.ch009
92. Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study. *Discover Artificial Intelligence*, 4(1), 2. https://doi.org/10.1007/s44163-023-00079-z
93. Wang, T., Lund, B. D., Marengo, A., Pagano, A., Mannuru, N. R., Teel, Z. A., & Pange, J. (2023). Exploring the potential impact of artificial intelligence (AI) on international students in higher education: Generative AI, chatbots, analytics, and international student success. *Applied Sciences*, 13(11), 6716. https://doi.org/10.3390/app13116716
94. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). https://doi.org/10.18653/v1/2020.emnlp-demos.6
95. Xu, W., Wu, Y., & Ouyang, F. (2023). Multimodal learning analytics of collaborative patterns during pair programming in higher education. *International Journal of Educational Technology in Higher Education*, 20(1), 8. https://doi.org/10.1186/s41239-022-00377-z
96. Zeineddine, H., Braendle, U., & Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. *Computers & Electrical Engineering*, 89, 106903. https://doi.org/10.1016/j.compeleceng.2020.106903
97. Zhao, Y., Guo, Y., & Wang, X. (2025). Hybrid LSTM–Transformer Architecture with Multi-Scale Feature Fusion for High-Accuracy Gold Futures Price Forecasting. *Mathematics*, 13(10), 1551. https://doi.org/10.3390/math13101551

Publishing and Ethical Statements / بيانات النشر والالتزام الأخلاقي

N		Publication Data in English	بيانات النشر بالعربية		م
1	Authors' Contributions:	First Author: Design, methodology, data collection, analysis, and drafting. Second Author: Supervision, summarizing, development, and final review for publication.	الباحث الأول: التصميم، المنهجية، جمع وتحليل البيانات، وكتابة المسودة. الباحث الثاني: الإشراف العلمي، التلخيص، التطوير، والمراجعة النهائية للنشر.	مساهمة الباحثين:	1
2	Conflict	No conflicts of interest.	لا يوجد تضارب مصالح.	تضارب المصالح:	2
3	Funding	Self-funded (No external grant).	تمويل ذاتي (لا يوجد دعم خارجي).	التمويل:	3
4	Copyright Licensed under:	(CC BY-NC-ND)		حقوق النشر مرخص بموجب:	4
5	ReviewProcess:	Double-blind peer review.	تحكيم مزدوج التعمية.	آلية التحكيم:	5
6	Plagiarism Check:	Verified via (iThenticate)	تم الفحص عبر	فحص الانتحال:	6
7	Data Availability:	Available upon request.	متاحة عند الطلب.	إتاحة البيانات:	7