e-REVISTA MULTIDISCIPLINARIA DEL SABER
ISSN: 2960-2467

https://doi.org/10.61286/e-rms.v4i.381

Original Article

Explainable AI for assessment design optimization in undergraduate education

Inteligencia artificial explicable para la optimización del diseño de la evaluación en la educación universitaria

Sinan, Adnan Muhisn

College of Biotechnology, Al-Qasim Green University, Babylon, Iraq.

Abstract

Assessment design plays a pivotal role in shaping undergraduate learning outcomes, but often the choice of assessment structure and weighting is made following the pedagogical intuition rather than research data. Although machine learning has gained significant use in the higher education sector to forecast the performance of students, insufficient effort has gone into showing how the elements of assessment per se affect academic performance. This paper proposes an assessment-centric explainable artificial intelligence (XAI) framework for analyzing and optimizing assessment design in undergraduate education. Real course assessment data in the form of assessment weights, frequency and diversity machine learning models are trained to predict course outcomes and SHapley Additive exPlanations (SHAP) are used to measure the contribution of individual assessment components. Unlike existing student-centric explainability approaches, the proposed framework focuses on assessment structures, enabling transparent analysis of how design choices affect success and failure risk. According to the outcomes of the experiment, high stakes final exams are strongly related with the probability of failures, whereas the diversified strategies of continuous evaluation decrease the probability of failures. The outcomes provide a feasible information to the curriculum developers and academic decision-makers that could help them to redesign assessment based on evidence.

***Keywords**: explainable artificial intelligence, machine learning, education, assessment design, undergraduate education, learning analytics, curriculum optimization, educational data mining.*

Resumen

El diseño de la evaluación desempeña un papel fundamental en la configuración de los resultados de aprendizaje en la educación universitaria; sin embargo, la elección de la estructura y ponderación de las evaluaciones suele basarse en la intuición pedagógica más que en evidencia empírica. Aunque el aprendizaje automático ha sido ampliamente adoptado en la educación superior para predecir el rendimiento estudiantil, se ha prestado escasa atención a analizar cómo los propios componentes del diseño de la evaluación influyen en el desempeño académico. Este trabajo propone un marco de inteligencia artificial explicable (XAI) centrado en la evaluación para el análisis y la optimización del diseño de la evaluación en la educación universitaria. A partir de datos reales de cursos —incluyendo ponderaciones, frecuencia y diversidad de los componentes evaluativos se entrenan modelos de aprendizaje automático para predecir los resultados académicos, mientras que las explicaciones basadas en SHapley Additive exPlanations (SHAP) se emplean para cuantificar la contribución individual de cada componente de la evaluación. A diferencia de los enfoques existentes, predominantemente centrados en el estudiante, el marco propuesto se enfoca en las estructuras de evaluación, permitiendo un análisis transparente de cómo las decisiones de diseño se asocian con el éxito académico y el riesgo de fracaso. Los resultados experimentales indican que las evaluaciones finales de alto impacto se asocian fuertemente con una mayor probabilidad de fracaso, mientras que las estrategias diversificadas de evaluación continua reducen dicho riesgo. Estos hallazgos proporcionan información práctica y basada en evidencia para desarrolladores curriculares y responsables de la toma de decisiones académicas, apoyando el rediseño informado de los sistemas de evaluación.

Palabras Clave: artificial explicable, aprendizaje automático, educación, diseño de la evaluación, educación universitaria, analítica del aprendizaje, optimización curricular, minería de datos educativos.

Recibido/Received	08-02-2026	Aprobado/Approved	13-04-2026	Publicado/Published	16-04-2026

CC BY NC SA

Introduction

Assessment design is a central component of undergraduate education, shaping learning behaviors, student engagement, and academic outcomes. Decisions regarding the structure and weighting of assessment components such as final examinations, quizzes, assignments, and laboratory activities are typically informed by pedagogical experience, institutional norms, or accreditation requirements. Despite their importance, these decisions are often made without systematic, data-driven evidence regarding how assessment structures influence student success and failure. Recent assessment analytics studies focus on automation and feedback generation but have little support in assessing assessment design decisions in a way that is applicable and interpretable (Hooda et al., 2022; Martin et al., 2025).

The increased use of machine learning (ML) in the educational sector and higher education has resulted in significant advancements in student performance prediction, the identification of at-risk students, and aid in academic decision-making. Using the data of learning management systems, the results of assessments, and the indicators of engagement, ML-based methods have demonstrated great predictive quality in undergraduate settings (Akçapınar et al., 2019; Islam et al., 2025). Nevertheless, most of these studies have a student-centred approach, which puts emphasis on individual factors such as behaviour, demographics, or fairness, instead of addressing the structure of assessment design as such.

Educational research has long established that assessment structure plays a critical role in learning. Formative and continuous assessment strategies are associated with improved self-regulation, sustained engagement, and deeper learning, whereas higsh-stakes summative assessments have been linked to increased stress and performance volatility (Nicol & Macfarlane-Dick, 2006; Yorke, 2003). Even though these educational findings are widely established, past empirical research generally uses qualitative analysis or standard statistical methods and lacks scalable systems that can model nonlinear interactions between assessment components.

With the penetration of ML models into educational decision-making, transparency, trust, and accountability issues increase. Black-box predictive models will be limited to instructors and curriculum designers who must have interpretable evidence to support assessment decisions. An important response to this challenge has been Explainable Artificial Intelligence (XAI) which allows the predictions of a model to be explained in a way that is human understandable (Doshi-Velez & Kim, 2017). Among XAI techniques, SHapley Additive exPlanations (SHAP) provide a theoretically grounded framework for attributing predictive outcomes to individual features (Lundberg & Lee, 2017).

Recent work increasingly integrates XAI into educational data mining and learning analytics to explain student performance, fairness, and feature importance (Efendi, 2025; Kesgin et al., 2025; Ofori & Dake, 2025). Systematic reviews confirm that SHAP and related techniques are now widely used to interpret machine learning models in educational contexts (Altukhi & Pradhan, 2024; Pachouly & Bormane, 2025). Nevertheless, these efforts remain predominantly student-centric, focusing on why individual students succeed or fail, rather than explaining how assessment design decisions influence outcomes at the course or curriculum level.

Research on learning analytics demands more methods that go beyond descriptive dashboards to actionable, decision oriented analytics that can be used in curriculum and assessment design (Kaliisa et al., 2024). Although this call is made, there are still explainable frameworks that explicitly quantify the impact of assessment structures including assessment weighting, balance, and diversity which are hardly explored.

ML techniques have been widely applied in educational settings to predict student performance, dropout risk, and academic success. Early studies focused on supervised learning models using demographic, behavioral, and academic features to identify at-risk students (Baker & Yacef, 2009; Romero & Ventura, 2010). More recent works have employed ensemble models and deep learning architectures to improve predictive accuracy using learning management system (LMS) data (Akçapınar et al., 2019) .

Due to the recent attention of researchers to the field of machine learning and its practical use in undergraduate education to predict academic performance, persistence, and graduation outcomes. Osunbunmi et al. (2025) report on research exploring how ML models and XAI assessments can improve predictions of undergraduate engineering student retention, they employ SHAP and LIME, to analyze engineering students' persistence to graduation using a large longitudinal dataset. Their findings identify early GPA, standardized test scores, and performance in foundational STEM courses as key predictors, extending persistence theory through data-driven modeling; however, the explanations remain focused on student-level attributes rather than instructional or assessment design factors.

Assessment is the process whereby it involves collecting and evaluating information in accordance to certain criteria in order to arrive at a judgment. Assessment and feedback are an important aspect of the learning process and their correlation with learning, teaching and curriculum has never been a trivial factor towards successful learning outcomes and enhancing student satisfaction (Hooda et al., 2022). Furthermore, assessment play a crucial role in the teaching and learning process by facilitating learning, promoting engagement, and enhancing student participation (Martin et al., 2025).

Numerous studies have explored the relationship between assessment strategies and student learning outcomes. Educational research has shown that continuous assessment and formative feedback can improve student engagement and reduce failure rates(Hooda et al., 2022; Nicol & Macfarlane-Dick, 2006). Feedback provided to students can serve as a diagnostic tool for instructors, helping them tailor their teaching strategies to better meet the needs of their students (Yorke, 2003) . However, most assessment studies rely on statistical analysis or qualitative evaluation and lack widely applicable, data driven methodologies. Furthermore, these studies do not leverage ML models capable of capturing complex nonlinear relationships between assessment components and outcomes.

Since ML is used in education, concerns regarding transparency, integrity, and trust in automatic decision making have grown. Explainable Artificial Intelligence (XAI) has become one of the methods of mitigating such concerns by providing explanations that can be understood and assessed by stakeholders to interpret AI-aware decisions. In educational research, XAI has been primarily applied to analyzing student outcomes, such as explaining individual risks of failure or dropout (Kizilcec, 2016; Lundberg & Lee, 2017). The integration of more advanced AI techniques and XAI, holds promise for the future of automated assessment and detect important behavior and academic traits affecting such predictions (Holstein et al., 2019). Recent studies highlight the use of XAI in educational analytics, Latif et al. (2023) propose a modified machine learning model for student performance prediction and identification of students at risk, though explanations remain limited to behavioral indicators rather than assessment design. Furthermore, Leichtmann et al. (2023) showed that users who received explanations of AI predictions were better able to understand and trust the system.

XAI tools can provide insights into how AI systems reach their conclusions, enabling educators to critically evaluate and trust these technologies to provides good guidance for educators and IT solution architects in the context of education (Khosravi et al., 2022). Additional reviews, by Fiok et al. (2022) and Kalasampath et al. (2025) also emphasize the relevance of transparency and recall SHAP and LIME as common techniques, though also note the weaknesses of education-focused frameworks of analysis. Despite the growing interest in XAI within education, previous research has still focused primarily on the student, providing limited insight into structural decisions such as assessment composition and weighting. This gap motivates the present study, which shifts the focus of explainability from student outcomes to assessment components, enabling transparent and actionable insights for assessment design optimization and curriculum-level decision-making.

Recently, XAI has been used to add value to transparency, fairness, and trust to educational ML systems. Lünich and Keller (2024) empirically examine the effect of model interpretability on student perceptions of fairness and causality in academic performance prediction through a large-scale factorial survey and show that simpler, more interpretable decision-tree models have a stronger effect on

perceived fairness, whereas essential predictive accuracy has less effect. Their results highlight the moral significance of explainability in learning analytics.

Tariq et al. (2025) develop an explainable framework for predicting student stress using psychosocial and survey-based variables, with SHAP analysis highlighting factors such as sleep quality and teacher–student relationships; while valuable for institutional support strategies, the approach does not address academic or assessment design considerations. Finally, SANFO (2025) applies supervised learning and SHAP to predict student learning outcomes using large-scale educational data, identifying community involvement, school infrastructure, and teacher experience as dominant predictors, but focusing primarily on contextual and socio-institutional factors rather than assessment-related variables.

Islam et al. (2025), integrate a variety of ML classifiers with XAI methods, such as SHAP and LIME, in order to predict undergraduate academic performance, with the highest accuracy on XGBoost and most significantly important predictors, both individual and contextual. Nonetheless, the suggested explainability is student centered and does not touch on assessment design and curriculum structure. However, Ofori and Dake (2025), suggest an interpretable hybrid model of LSTM-Transformer type predicting the performance of students, which employs SHAP and attention mechanisms to reveal significant behavioral and demographic predictors. Though deep learning is made more interpretable, the composition of assessment and grading structures are not taken into consideration with this framework.

Kesgin et al. (2025) advance fairness-aware student performance prediction by integrating SHAP and debiasing techniques across multiple ML models, while Efendi (2025) employs SHAP-driven feature engineering to improve adaptability classification in online learning. Despite their contributions to responsible and interpretable educational AI, both studies remain learner-centric and do not address assessment design or course-level structural decisions. Finally, Altukhi and Pradhan (2024) systematically review XAI in education using the PRISMA methodology, identifying key definitional gaps and ethical, technical, trust, and policy challenges. The review emphasizes the need for application specific explainability frameworks but does not address assessment design or curriculum level decision making, reinforcing the research gap addressed in the present study.

All of these studies point to the increasing maturity of XAI techniques in education to improve interpretability, fairness, and decision support. Nonetheless, these studies are still largely student or situation centric with little understanding of the way design components of assessments affect the outcome in academic achievement, thus supporting the drive behind assessment centric explainable frameworks. Irrespective of these developments, no current literature exists that utilizes explainable AI to examine assessment design as a structural predisposition of undergraduate achievement.

This paper argues that explainability in educational ML should extend beyond the interpretation of individual student predictions to support assessment-focused academic decisions. To develop a curriculum based on evidence, one should have a understanding the contribution of assessment elements to academic achievements and failures. To that end, the study presents an explainable AI framework, combining predictive modeling with SHAP-based interpretations to examine the effect of assessment design on undergraduate academic outcomes. By shifting the locus of explainability from learner characteristics to assessment structures, the framework provides transparent and practically relevant insights to improve the design of assessment in the real undergraduate environment.

This paper presents a novel XAI driven framework that interprets undergraduate academic outcomes through the lens of assessment design elements, rather than student-specific attributes. By integrating machine learning–based prediction with SHAP, the proposed approach delivers transparent and actionable insights that support evidence-based evaluation of assessment structures and informed curriculum redesign. The framework relies exclusively on commonly available institutional assessment data, enabling practical deployment in real undergraduate settings without requiring sensitive personal or demographic information. In doing so, this work advances the role of explainable AI in education beyond

predictive transparency toward curriculum-level decision support, thereby bridging educational theory and data-driven assessment optimization.

Materials and methods

1. Research design

This study employs an assessment-centric explainable machine learning design to examine the impact of assessment frameworks on undergraduate academic achievements. The suggested methodology prioritizes the explanatory study of assessment design elements as decision variables, unlike traditional educational machine learning studies that focus on predicting student performance. The research design follows a two-stage process: firstly, predictive modeling of course outcomes using assessment-related features. Secondly, explainability driven interpretation of model predictions to quantify the contribution of individual assessment components.

The methodological approach relies upon the famous assessment theories and the constructive alignment, in particular, which is concerned with the consistency of the learning outcomes, the instructional activities, and the assessment designs, as well as formative assessment theory, which highlights the role of continuous assessment in supporting self-regulated learning (Nicol & Macfarlane-Dick, 2006). The study is explanatory in nature and focuses on identifying **associative relationships** between assessment design choices and academic outcomes rather than establishing causal inference.

2. Dataset description

• **Data source:** The dataset consists of anonymized academic records collected from undergraduate courses offered by a public university in Iraq. The institutional academic records and the learning management system (LMS) were used to gather the data over the course of several academic semesters. The dataset illustreated in **Table 1** consists of assessment scores, assessment weighting plans and final course results. In order to comply with ethics, no sensitive personal characteristics (e.g., ethnicity, health data or financial records) were mentioned. Any data analysis was conducted in aggregate form and was not used in any other way other than in research.

The courses that were analyzed are part of only one environment, the College of Biotechnology of one of the Iraqi state universities, which guarantees the discipline consistency. The courses were picked on the basis of having complete assessment structure data and final outcome records. The instances of student-course where the final outcome data was not complete were not included in the analysis. Missing data in the elements of assessment were filled with median imputation. Students who officially dropped a course before the end of assessment were eliminated to avoid the distortion of the pass/fail results.

Table 1. *Summary of the dataset*

Attribute	Description
Institution	College of Biotechnology, public university (Iraq)
Academic level	Undergraduate
Number of students	360
Number of courses	16
Semesters covered	4 semesters
Assessment types	Exams, quizzes, assignments, laboratories
Outcome variable	Pass/Fail or Grade (A–F)
Data source	Institutional records and LMS
Sensitive personal data	None

• **Unit of analysis:** The unit of analysis is defined as a student–course instance, representing a student's complete assessment record within a specific course. While individual students may appear in multiple courses, the analysis focuses on course-level assessment structures, which remain consistent across students enrolled in the same course. Potential intra student dependency is acknowledged; however, it does not invalidate the assessment centric explanatory objective of the study, as the emphasis lies on assessment design rather than individual learner profiling.

3. Feature engineering

Feature engineering was guided by pedagogical theory and prior assessment analytics research. Because performance aggregates are partially derived from assessment outcomes, they were included solely as control variables to stabilize prediction and were excluded from all explainability analyses. To assess potential label leakage,

an ablation study was conducted comparing models with assessment-structure features only, performance aggregates only, and their combination. Features were categorized into three groups:

• **Assessment structure features:**

Final examination weight: Proportion of the final exam score relative to the total course grade.

Continuous assessment ratio: Aggregate weight of quizzes, assignments, and laboratory work relative to summative assessment.

Assessment diversity index: A measure capturing the distribution of assessment weights across different assessment types. This index is computed using an entropy-based formulation derived from information theory, where higher entropy indicates greater diversity in assessment components.

The inclusion of these features operationalizes formative and summative assessment principles and aligns with prior educational research emphasizing balanced assessment design (Yorke, 2003) .

• **Performance aggregate features:**

Aggregated performance indicators (e.g., average quiz score, assignment score and laboratory score) were added to control baseline student performance. In the explainability analysis, these features were not interpreted, as it would confound effects of student performance with the effects of assessment design.

• **Contextual features:**

Basic contextual attributes such as course credit value and semester were included to capture structural differences across courses. The entropy-based one is based on the well-known principles of information-theoretic and has already been used to describe assessment diversity in education (Yorke, 2003).

$$ADI = \frac{-\sum_{i=1}^{K} w_i \log(w_i)}{\log(K)}$$

In Eq. (1), wi denotes the normalized weight of assessment component i, and K represents the total number of assessment components in the course. Each assessment weight is first normalized such that $\sum_{i=1}^{k} \mathrm{wi} = 1$. The division by log(K) ensures that the Assessment Diversity Index (ADI) is bounded between 0 and 1, where values close to 0 indicate low diversity (dominance of a single assessment type), and values close to 1 indicate a balanced distribution across multiple assessment components. For example, a course assessed solely through a single final examination yields an ADI value of 0, reflecting minimal assessment diversity. In contrast, a course with four equally weighted assessment components (e.g., exam, quizzes, assignments, and laboratory work) achieves an ADI value of 1, indicating maximal diversity.

Course credit value and semester were included as contextual controls. Instructor and cohort effects could not be explicitly modeled and are acknowledged as potential confounders. Furthermore, because assessment weights are compositional and sum to one, collinearity is possible. Tree-based models are less sensitive to such constraints; nevertheless, future work may explore log-ratio transformations for enhanced interpretability (Pawlowsky-Glahn et al., 2015).

4. Data preprocessing

Data preprocessing included normalization of numerical features, handling of missing values using median imputation, and stratified splitting of the dataset into training and testing subsets. The stratification ensured proportional representation of pass and fail outcomes in both subsets. Class imbalance was assessed and found to be within acceptable limits; therefore, no resampling techniques were applied.

In addition to stratified five-fold cross-validation at the instance level, a grouped cross-validation experiment was conducted by holding out entire courses during training to assess generalization of assessment design insights beyond observed courses.

5. Machine learning models

The tree-based ensemble models were chosen because of its high level of empirical effectiveness in structured tabular educational data and because of its knack to broaden the intricate non-linear associations and

mixes of features without the need to carry out a significantly high level of feature transformation. In particular, the models of the random forest and the gradient boosting (XGBoost) were used, because these algorithms are best applied to medium-size datasets and heterogeneous feature space common in institutional assessment data. Ensemble methods decrease the variance and enhance the generalization because they combine numerous decision trees and hence increase the predictive stability. In addition, tree-based architectures can be used with TreeSHAP, which allows to perform explainability analysis effectively and theoretically, which is computationally efficient. The models have been popular in the academic outcome prediction of educational data mining and learning analytics (Akçapınar et al., 2019; Romero & Ventura, 2010).

To provide methodological contrast and avoid exclusive reliance on non-linear ensemble approaches, a logistic regression baseline model was implemented. This baseline facilitates comparison between linear and non-linear decision boundaries and serves as a transparent reference point, mitigating concerns regarding over-dependence on complex black-box models.

6. Explainability framework

SHapley Additive exPlanations (SHAP) are employed to interpret model predictions. SHAP values provide a theoretically grounded method to quantify the marginal contribution of each assessment component to the predicted outcome. SHAP has been increasingly adopted in educational research due to its ability to enhancemodel interpretability and support data-driven decision-making (Efendi, 2025; Guan et al., 2025). TreeSHAP was employed using the training dataset as the background distribution. Stability was assessed via cross-validation and bootstrapped resampling, yielding consistent feature rankings. The explainability analysis was conducted at multiple levels:

Global explanations to identify assessment components with the strongest overall influence on outcomes.

Local explanations to analyze how specific assessment configurations contribute to success or failure scenarios.

Scenario-based explanations to simulate hypothetical changes in assessment design and observe their impact on predicted outcomes.

While SHAP provides consistent and model-agnostic explanations, its limitations—such as sensitivity to correlated features and post-hoc interpretation—are acknowledged. These limitations were mitigated by careful feature grouping and by focusing interpretation on assessment structure variables.

7. Evaluation metrics

Predictive performance was evaluated using standard classification metrics, including accuracy, F1-score, and area under the receiver operating characteristic curve (AUC). Explainability quality was evaluated qualitatively based on interpretability, actionability, and consistency, following established guidelines for explainable AI in education (Doshi-Velez & Kim, 2017; Khosravi et al., 2022) .

Interpretability was assessed by the clarity of feature contributions, while actionability was evaluated based on whether explanations could reasonably inform assessment redesign decisions.

8. Ethical considerations

The framework proposed is aimed at enabling the human-in-the-loop decision-making instead of automated intervention. Explanations are supposed to help the instructors and curriculum designers and not to compare or brand certain students. Sensitivity of the attribute's exclusion and emphasis on assessment structures lead to ethical adherence and minimize the risks of bias or stigmatization.

Results

Predictive performance evaluation

The predictive performance of the proposed models was evaluated using accuracy, F1-score, and area under the ROC curve (AUC). To ensure robustness, all models were assessed using stratified five-fold cross-validation, and results are reported as mean ± standard deviation across folds.

Tree-based ensemble models demonstrated consistent and stable performance across all metrics. In particular, the Gradient Boosting model achieved the highest predictive performance, with an average

accuracy of 0.83 ± 0.03, F1-score of 0.81 ± 0.04, and AUC of 0.86 ± 0.02. The random forest model achieved similar results, confirming that ensemble methods are suitable for tabular educational data. The obtained AUC values are comparable to those reported in recent ML-based educational research (Islam et al., 2025; Osunbunmi et al., 2025) .

To contextualize these results, a logistic regression baseline model was implemented. While the baseline achieved reasonable performance (accuracy: 0.72 ± 0.05, AUC: 0.78 ± 0.04), it was consistently outperformed by tree-based models across all folds. A paired t-test confirmed that the performance improvements achieved by ensemble models were statistically significant ($p < 0.01$), justifying the use of non-linear models for subsequent explainability analysis.

Table 2 predictive performance on stratified fivefold cross validation of ML models. The findings are presented in terms of mean± standard deviation. Ensemble models that utilize tree-based always perform better than the logistic regression baseline and Gradient Boosting is best overall. The levels of those performances are similar or higher than those achieved in the previous educational prediction experiments based on the application of ML approaches, which confirms the validity of the experimental design.

Table 2. Comparison of ML Models for Course Outcome Prediction

Model	Description	Precision	Recall	F1-Score	AUC
Logistic Regression (Baseline)	Linear baseline model used to contextualize the performance of non-linear approaches	0.72 ± 0.05	0.75 ± 0.06	0.73 ± 0.05	**0.78 ± 0.04**
Random Forest	Ensemble tree-based model capturing non-linear relationships in tabular educational data	0.80 ± 0.04	0.82 ± 0.04	0.80 ± 0.04	**0.85 ± 0.03**
Gradient Boosting	Sequential ensemble model optimizing predictive accuracy through iterative error correction	**0.82 ± 0.03**	**0.83 ± 0.03**	**0.81 ± 0.04**	**0.86 ± 0.02**

Global explainability analysis

Global explainability was conducted using SHAP to identify assessment design features with the greatest overall influence on course outcomes. Feature importance was computed by aggregating absolute SHAP values across all instances and folds. Furthermore, Assessment diversity, quantified using the Assessment Diversity Index defined in Eq. (1), emerged as a consistently influential factor in the explainability analysis.

The findings suggest that final examination weight, ratio assessment continuity and assessment diversity index are the predictors of course success and failure that were consistent. Aggregate features of performance showed predictive relevance, and were not to be interpreted to maintain the assessment-centric focus of the analysis.

It is important to note that SHAP-based explanations reflect associative relationships learned by the model and do not imply causal effects. Nevertheless, these associations provide valuable insight into how assessment structures systematically relate to academic outcomes.

SHAP dependence and threshold analysis

SHAP dependence plots were employed to examine non-linear relationships between key assessment features and predicted failure risk. The analysis reveals a pronounced increase in predicted failure probability as the final examination weight exceeds approximately 45–50% of the total course grade. Rather than asserting a definitive threshold, these results suggest a potential risk zone associated with high-stakes summative assessment.

On the other hand, increased continuous assessment ratios were related with lower risk of failure, especially when diversified assessment format was in use. Assessment diversity showed significant

monotonic correlation with the probability of success, such that a mixed evaluation strategy can protect students against failure. These findings align with established assessment theory emphasizing the pedagogical benefits of formative and diversified assessment practices, while remaining grounded in associative explanatory evidence.

Course-level explainability analysis

SHAP data were grouped on courses so as to allow comparison of the effects of assessment design across courses. This discussion showed that there are great differences in the effects of assessment formats.

Courses that had a standardized assessment plan had a lower chance of failure prediction compared to courses that had a higher number of final examinations. Notably, the trends were replicated in each and every student cohort and they reflected the structural effect of the assessment design without any regard to the individual learner characteristics. This course-level elucidation offers practical insights for curriculum committees and educators aiming to analyze and reformulate evaluation systems.

Scenario-based explainability analysis

In order to understand the practical aspects of explainability, scenario-based SHAP analysis was performed through the simulation of alternative assessment settings. In particular, weight of final examination was decreasing step by step with redistribution of weight to quizzes and assignments.

The simulated cases suggest that the assessment weight can be redistributed to continuous components to significantly lower the risk of predicted failure. For example, reducing final exam weight from 60% to 40% was associated with an average reduction in predicted failure risk of approximately 20–25%. These findings demonstrate that explainable models can be applied to test out hypothetical assessment redesigns before they are implemented.

Although such situations do not institute causal outcomes, they offer clear evidence-based direction on assessment planning. In addition, all scenario-based simulations maintained the compositional constraint (weights adding up to one) and institutional policy constraints that were seen in the dataset.

Explainability robustness and stability

The robustness of explainability results was evaluated by examining the stability of SHAP feature rankings across cross-validation folds. Spearman rank correlation coefficients were computed for global feature importance rankings, yielding an average correlation of 0.89, indicating high stability. While grouped cross-validation resulted in a modest reduction in predictive performance, the relative importance and ranking of assessment design features remained stable.

Moreover, the difference in SHAP values of the key assessment features was low at all the stages of the study, which confirms the reliability of the interpretability analysis. Such findings show that the significance of assessment design characteristics observed is not caused by data segmentation or variant

The experimental results confirm that the features of assessment design are not only predictive of the outcomes of a undergraduate course, but also can be explained in a stable and interpretable way using SHAP-based analysis. Specifically, high-stakes final tests demonstrate a reliable correlation with the high risk of failure, and the framework of diversified and continuous assessment is linked with better academic performance. These findings are consistent with the known pedagogical concepts of the importance of formative assessment and the balanced grading models, and they present quantitative and explicable findings based on actual institutional data.

Notably, the suggested framework is not confined to predictive performance but will provide multi-level explainability, such as, global feature importance, course-level aggregation, and scenario-based analysis. This interpretability by layers allows the stakeholders to shift out of statistical relationships to

actionable design implications, and the methodological transparency. Any interpretations are presented in a form of association and not causality, as is also in line with responsible XAI practices in educational analytics.

Table 3 places the current work in the context of the larger explainable AI application in education. As demonstrated, previous studies have mainly concentrated on student-level forecasting, behavioral analytics, and model output transparency. Majority of current methods elucidate personal learner traits, demographic attributes or activity-related aspects in learning management systems. Even powerful models like XAI-ED are still rather theoretical with no empirical analysis of assessment systems. Contrarily, the proposed work is the first to use explainable machine learning in areas of assessment design, where the unit of explanation is not an individual learner but the course-level grading framework. This redefining of explainability as curriculum formulation is a substantive methodological and conceptual development.

Table 3. *Comparison of XAI applications in education*

Study (Ref.)	Educational Task	ML Model	XAI Technique	Explained Entity	Level of Analysis	Limitation
Kizilcec (2016)	Learning analytics transparency	Statistical / ML models	Transparency indicators	Student behavior	Individual	No structural or design-level insight
Lundberg and Lee (2017)	Model interpretation (general)	Tree-based models	SHAP	Model features	Individual	Not education-specific
Holstein et al. (2019)	Student performance support	ML models	SHAP / LIME	Student features	Individual	Focused on learner attributes
Latif et al. (2023)	At-risk student detection	Ensemble ML	Feature importance	LMS activity patterns	Individual	No assessment design analysis
Leichtmann et al. (2023)	Trust in AI systems	Rule-based & ML	Visual explanations	Decision rationale	System-level	Not education-specific
Khosravi et al. (2022)	XAI framework for education	Conceptual	XAI-ED framework	Educational decisions	Conceptual	No empirical assessment analytics
Fiok et al. (2022)	XAI for education & training	Survey-based	Multiple XAI methods	Educational models	Conceptual	Lacks quantitative validation
Kalasampath et al. (2025)	XAI across domains	Various	SHAP, LIME	Model predictions	Cross-domain	Education treated marginally
Proposed Work	**Assessment design optimization**	**Tree-based ML**	**SHAP**	**Assessment components**	**Course / curriculum**	**None (assessment-centric)**

The study provides a solution to an important gap in educational data mining and learning analytics by focusing on explainability in terms of composition of assessment and not the characteristics of students.The findings substantiate that explainable artificial intelligence is capable of serving a dual function: enhancing predictive transparency while simultaneously facilitating evidence-based approaches to curriculum refinement. As such, the proposed framework makes a meaningful contribution to the integration of educational theory with data-informed decision-making processes, yielding a methodologically rigorous and ethically conscientious instrument for academic administrators and curriculum designers engaged in the systematic evaluation and reconceptualization of assessment practices.

Discussion

The current study presents a significant advancement in the application of Explainable Artificial Intelligence (XAI) within the educational domain. By shifting the focus from individual student traits to the structural components of assessment design, this research addresses a critical gap in educational data mining. The results demonstrate that tree-based ensemble models, particularly Gradient Boosting, provide a robust and statistically superior framework for predicting academic outcomes compared to traditional linear baselines. This superiority, confirmed by a paired t-test ($p < 0.01$), justifies the shift toward non-linear models when paired with sophisticated interpretability tools like SHAP (Lundberg & Lee, 2017).

The predictive performance achieved in this study (Gradient Boosting AUC: 0.86 \pm 0.02) is consistent with, and in some cases exceeds, recent benchmarks in the field. For instance, the results align with findings by Islam et al. (2025) and Osunbunmi et al. (2025), who utilized ensemble methods to predict student persistence and performance. The stability of these metrics across stratified five-fold cross-validation underscores the reliability of using machine learning for institutional decision-making. Unlike earlier studies that relied on simpler statistical models (Baker & Yacef, 2009; Romero & Ventura, 2010), the current approach leverages the high capacity of ensemble learning while maintaining transparency through XAI.

A primary contribution of this work is the identification of specific assessment features that influence failure risk. Global explainability analysis revealed that final examination weight, the continuous assessment ratio, and the Assessment Diversity Index are key predictors of course outcomes. Specifically, the SHAP dependence analysis identified a "risk zone" when the final examination weight exceeds 45–50% of the total grade. This finding provides empirical support for established pedagogical theories advocating for a move away from high-stakes summative evaluation (Yorke, 2003). The positive correlation between assessment diversity and success suggests that a multifaceted evaluation strategy acts as a protective factor for students. This aligns with the principles of formative assessment described by Nicol and Macfarlane-Dick (2006), where diverse feedback loops and varied task types support self-regulated learning. By quantifying these relationships, our model transitions from theoretical pedagogical advice to evidence-based curriculum design (Martin et al., 2025).

Unlike existing XAI-ED frameworks that often remain conceptual (Khosravi et al., 2022) or focus exclusively on individual student behavior (Kizilcec, 2016; Latif et al., 2023), the proposed framework offers multi-level insights. The course-level aggregation allows academic committees to identify systemic issues in grading structures regardless of student cohorts. Furthermore, the scenario-based analysis demonstrates the practical utility of SHAP for "what-if" simulations. The finding that redistributing 20% of exam weight to continuous assignments could reduce predicted failure risk by approximately 20–25% offers a powerful tool for proactive curriculum reform. This move toward "causability"—the human understanding of the "why" behind a model—is a crucial step for building trust in AI systems (Lünich & Keller, 2024; Doshi-Velez & Kim, 2017).

Table 3 highlights the novelty of this research relative to the broader literature. While studies by Akçapınar et al. (2019) and Tariq et al. (2025) have successfully implemented early-warning systems for at-risk students, they primarily utilize learner-centric data such as LMS logs or stress indicators. Similarly, recent work by Johora et al. (2025) and Ofori & Dake (2025) explores SHAP and LSTM-Transformer models for performance analysis but continues to focus on individual student attributes. In contrast, our work redefines the "explained entity" from the student to the assessment system itself. This shift is vital because, as noted by Kaliisa et al. (2024), many learning analytics dashboards have failed to impact achievement due to a lack of actionable, design-level insights. By focusing on the curriculum, we move away from "fixing the student" and toward "optimizing the environment," a perspective that is often missing in cross-domain XAI applications (Kalasampath et al., 2025; Kesgin et al., 2025).

The high stability of SHAP feature rankings (Spearman correlation of 0.89) ensures that the insights provided are not artifacts of data noise but reflect genuine structural patterns. This methodological rigor is essential when dealing with high-risk decisions in education (Leichtmann et al., 2023). However, it is imperative to maintain the distinction between association and causality. While SHAP values identify influential features, they do not guarantee that changing an assessment weight will result in a deterministic outcome, as educational environments are complex socio-technical systems (Altukhi & Pradhan, 2024; Efendi, 2025). Furthermore, the study respects the mathematical nature of grading systems by treating assessment weights as compositional data (Pawlowsky-Glahn et al., 2015), ensuring that simulations remain within institutional policy constraints. This ethical and grounded approach to AI integration is consistent with the latest trends in responsible XAI-ED (Fiok et al., 2022; Pachouly & Bormane, 2025).

In conclusion, the results substantiate that XAI can serve a dual function: enhancing predictive transparency and facilitating evidence-based curriculum refinement. By centering the analysis on assessment design, the proposed framework provides academic administrators with a rigorous tool to evaluate and reconceptualize evaluation practices. This study marks a transition from descriptive analytics to prescriptive, design-centric insights, bridging the gap between machine learning capabilities and educational theory (Hooda et al., 2022; Guan et al., 2025; Sanfo, 2025).

Final Considerations

The implementation of the proposed explainable framework marks a fundamental shift in how academic performance is analyzed, moving beyond the mere identification of at-risk students toward a proactive evaluation of pedagogical structures. By demonstrating that assessment design itself is a powerful predictor of academic success or failure, this study empowers curriculum designers with a quantitative tool to balance high-stakes examinations with continuous evaluation. This evidence-based approach ensures that institutional changes are not grounded in intuition but in reliable data that highlights the systemic influence of grading models on student outcomes.

Furthermore, the versatility of the multi-level explainability provided by SHAP allows for a transparent dialogue between data scientists and educators. The ability to simulate hypothetical assessment scenarios provides a risk-free environment to test curriculum reforms before they are formally adopted. This capability is essential for fostering a culture of continuous improvement within higher education, where the goal is to optimize the learning environment to be more inclusive and supportive, ultimately reducing failure rates through more diversified and balanced assessment strategies.

Finally, while the predictive power of ensemble models is undeniable, their true value in the educational sector lies in their interpretability. As AI continues to integrate into academic administration, maintaining a focus on responsible and transparent practices remains paramount. The findings of this study suggest that the future of learning analytics lies in the synergy between advanced machine learning and educational theory. Moving forward, applying this assessment-centric lens to broader datasets will be crucial for developing universal standards for effective and equitable curriculum design across diverse academic disciplines.

Acknowledgments

To our universities.

Conflicts of Interest

The authors declare no conflicts of interest.

Reference

Akçapınar, G., Altun, A., & Aşkar, P. (2019). Using learning analytics to develop early-warning system for at-risk students. International Journal of Educational Technology in Higher Education, 16(1), 1–20. https://doi.org/10.1186/s41239-019-0172-z

Altukhi, Z. M., & Pradhan, S. (2024). Systematic literature review: Explainable AI definitions and challenges in education. ICIS 2024 Proceedings.

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3–17. https://doi.org/10.5281/zenodo.3554657

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. https://doi.org/10.48550/arXiv.1702.08608

Efendi, Y. (2025). Machine learning-based classification of student adaptability in online learning with feature engineering. TIERS Information Technology Journal, 6(1), 129–143. https://doi.org/10.38043/tiers.v6i1.6806

Fiok, K., Farahani, F. V., Karwowski, W., & Ahram, T. (2022). Explainable artificial intelligence for education and training. The Journal of Defense Modeling and Simulation, 19(2), 133–144. https://doi.org/10.1177/15485129211028651

Guan, Y., Wang, F., & Song, S. (2025). Interpretable machine learning for academic performance prediction: A SHAP-based analysis of key influencing factors. Innovations in Education and Teaching International, 1–20. https://doi.org/10.1080/14703297.2025.2532050

Holstein, K., McLaren, B. M., & Aleven, V. (2019). Designing for complementarity: Teacher and student needs for orchestration support in AI-enhanced classrooms. International Conference on Artificial Intelligence in Education, 157–171. https://doi.org/10.1007/978-3-030-23204-7_14

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), Artículo 5215722. https://doi.org/10.1155/2022/5215722

Islam, M. M., Sojib, F. H., Mihad, M. F. H., Hasan, M., & Rahman, M. (2025). The integration of explainable AI in educational data mining for student academic performance prediction and support system. Telematics and Informatics Reports. https://doi.org/10.1016/j.teler.2025.100203

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, Artículo 100384. https://doi.org/10.1016/j.array.2025.100384

Kalasampath, K., Spoorthi, K. N., Sajeev, S., Kuppa, S. S., Ajay, K., & Angulakshmi, M. (2025). A literature review on applications of explainable artificial intelligence (XAI). IEEE Access. https://doi.org/10.1109/ACCESS.2025.3546681

Kaliisa, R., Misiejuk, K., López-Pernas, S., Khalil, M., & Saqr, M. (2024). Have learning analytics dashboards lived up to the hype? A systematic review of impact on students' achievement, motivation, participation and attitude. Proceedings of the 14th Learning Analytics and Knowledge Conference, 295–304. https://doi.org/10.1145/3636555.363688

Kesgin, K., Kiraz, S., Kosunalp, S., & Stoycheva, B. (2025). Beyond performance: Explaining and ensuring fairness in student academic performance prediction with machine learning. Applied Sciences, 15(15), 8409. https://doi.org/10.3390/app15158409

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence, 3, Artículo 100074. https://doi.org/10.1016/j.caeai.2022.100074

Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2390–2395. https://doi.org/10.1145/2858036.2858402

Latif, G., Abdelhamid, S. E., Fawagreh, K. S., Brahim, G. B., & Alghazo, R. (2023). Machine learning in higher education: Students' performance assessment considering online activity logs. IEEE Access, 11, 69586–69600. https://doi.org/10.1109/ACCESS.2023.3287972

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. Computers in Human Behavior, 139, Artículo 107539. https://doi.org/10.1016/j.chb.2022.107539

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

Lünich, M., & Keller, B. (2024). Explainable artificial intelligence for academic performance prediction: An experimental study on the impact of accuracy and simplicity of decision trees on causability and fairness perceptions. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 1031–1042. https://doi.org/10.1145/3630106.3658953

Martin, F., Kim, S., Bolliger, D. U., & DeLarm, J. (2025). Assessment types, strategies, and feedback in online higher education courses in the age of artificial intelligence: Perspectives of instructional designers. TechTrends, 1–17. https://doi.org/10.1007/s11528-025-01115-8

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. Studies in Higher Education, 31(2), 199–218. https://doi.org/10.1080/03075070600572090

Ofori, E., & Dake, D. K. (2025). Explainable artificial intelligence in LSTM transformer models for student performance analysis. Discover Computing, 28(1), Artículo 313. https://doi.org/10.1007/s10791-025-09814-9

Osunbunmi, I., Feyijimi, T., Cutler, S., Brijmohan, Y., Arinze, L., Dansu, V., Bamidele, B., Wu, J., & Rabb, R. (2025). Artificial intelligence in engineering education research: Using machine learning models to predict undergraduate engineering students' persistence to graduation. Journal of Engineering Education, 114(4), Artículo e70034. https://doi.org/10.1002/jee.70034

Pachouly, S., & Bormane, D. S. (2025). Explainable artificial intelligence in education: Transforming teaching and learning-a review. TPM–Testing, Psychometrics, Methodology in Applied Psychology, 32(S8), 1571–1584. https://doi.org/10.5281/zenodo.17866084

Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). Modeling and analysis of compositional data. John Wiley & Sons.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Sanfo, J.-B. M. B. (2025). Application of explainable artificial intelligence approach to predict student learning outcomes. Journal of Computational Social Science, 8(1), Artículo 9. https://doi.org/10.1007/s42001-024-00344-w

Tariq, R., Orozco-del-Castillo, M. G., Zamir, M. T., Ramírez-Montoya, M. S., & Wilberforce, T. (2025). Explainable artificial intelligence for predictive modeling of student stress in higher education. Scientific Reports, 15(1), Artículo 38375. https://doi.org/10.1038/s41598-025-22171-3

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. Higher Education, 45(4), 477–501. https://doi.org/10.1023/A:1023967026413