IJISEM
INTERNATIONAL JOURNAL OF INNOVATIONS IN SCIENCE ENGINEERING AND MANAGEMENT

OPEN ACCESS

Volume: 5

Issue: Special 1

Month: May

Year: 2026

ISSN: 2583-7117

Published: 09.05.2026

Citation:

Dr. Rakesh Kumar Pathak, Dr. Prakash Upadhyay "XAI-Based Evaluation Model for Navigating Academic Ambiguity" International Journal of Innovations in Science Engineering and Management, vol. 5, no. S1, 2026, pp. 53-58.

DOI:

10.69968/ijisem.2026v5Si153-58

XAI-Based Evaluation Model for Navigating Academic Ambiguity

Dr. Rakesh Kumar Pathak[1]

[1]*Assistant Professor, School of Computer Science, Xavier University, Patna*

Dr. Prakash Upadhyay[2]

[2]*Assistant Professor, School of Computer Science, Xavier University, Patna*

Abstract

Academic evaluation often does not reflect student's abilities accurately. There are sufficient testaments available which proves that student who did not scored well in their academic examinations have done exceptionally well in competitive exams have done very well in their life. Academic evaluation often becomes unclear because instructions are interpreted differently by teachers, raters score inconsistently, and automated grading systems are not transparent. Till date our assessment modules are largely based on memory-based evaluation and not on ability-based assessment. This paper suggests using an Explainable AI (XAI)–based evaluation model to reduce this ambiguity by:

1) Making automated and mixed (human + AI) scoring more transparent,

2) Showing clear explanations for each feature and each student's score, and

3) Providing measurable data that can help improve rubrics.

We include a synthetic-data experiment to show how an easy-to-understand model and XAI methods can

A. Highlight the main factors affecting grades and

B. Detect where strict grading or peer-review differences create ambiguity.

The findings show that XAI can offer practical insights that make academic assessment clearer, fairer, and more trustworthy. The paper ends with key suggestions and areas for future research

Keywords: Explainable AI, assessment clarity, rubric interpretation, fairness, educational evaluation.

INTRODUCTION

Artificial Intelligence (AI) has emerged as a vibrant and dynamic subject of study due to recent technological advancements. These days, AI systems are utilized in a wide range of contexts outside of research facilities, significantly influencing our day-to-day existence. By improving predicted performance through intricate models and algorithms, these systems can amplify, augment, and improve human performance [1] [2] [3]. However, AI systems that has black-box models, offer opaque decision-making, are the result of a strong concentration on prediction accuracy. [2]. In order to get around these challenges,

a lot of work has been done recently to develop explainable systems, which try to make AI systems and their results comprehensible to humans [4] [5][6]. The traditional AI models are often referred to as black box because what happens inside the model to arrive at any decision or conclusion, remains completely hidden

AI is now a crucial component of technology use. The European Union's General Data Protection Regulation (GDPR) commission has established users' rights to privacy, data protection, and algorithm transparency explanations for AI systems. As a result, all AI systems are supposed to be comprehensible [7]. Systems that use AI reasoning to forecast and make judgments are known as "black boxes." However, because it is a "black box," it is next to impossible for humans to comprehend how the system arrived at a conclusion or a forecast, let alone the degree of data confidentiality and safety of such systems [8]. According to the European Union's GDPR commission, this is a crucial issue [6]. The rationale is that industries like self-driving cars, robotic assistants, and tailored medicine cannot take the chance of making poor choices [9]. The authors of [9] have highlighted a number of "black box" issues that enable people to comprehend the reasoning behind a decision or prediction, including transparency, result explanation issues, and model inspection issues. In order to understand the internal process, explain the result, and make it more transparent using terminology and visuals displayed on the user interface (UI), XAI has emerged to our rescue and is there to open the "black box.

The goal of assessment in education is to properly and consistently gauge students' learning. However, ambiguity created by subjective rubrics, uneven rater assessments, and opaque automated scoring pipelines compromises the validity and confidence of evaluation results. In education, where stakeholders (students, teachers, and administrators) need practical, understandable explanations for scores, Explainable AI (XAI) provides techniques that make model judgments transparent and interpretable. This paper develops a practical XAI-based evaluation model intended to (i) identify sources of ambiguity, (ii) provide transparent instance-level and global explanations, and (iii) produce metrics that guide rubric refinement and rater calibration.

Review of literature

Ambiguity in academic evaluation stems from a combination of linguistic variables, subjective assessment standards, and the structural difficulties of educational institutions. It frequently results in inconsistent grading, unclear feedback, and challenges evaluating the caliber of proposals, instructors, or students. The ambiguity in assessment especially more evident when a group of students are evaluated by a group of evaluators belonging to different backgrounds and when their assessment styles are not rationalized in terms of their methods of awarding grades to the students. It has been quite evident that there are few evaluators who have the tendency being more strict than the rest or who are more lenient than the others. The authors of [10][11], addresses the ongoing confusion between the terms interpretability and explainability in machine learning. The authors propose a unified classification framework to help researchers and practitioners select and evaluate explainable AI (XAI) methods more effectively. The authors distinguish the two concepts based on how they achieve transparency:

Interpretability: Describes a model's innate ability to be comprehended by a person. It is tied to the model's structure (e.g., shallow decision trees or linear models).

Explainability: Involves offering understandable reasoning for decisions made by a model, frequently employing external, post hoc methods to clarify "black-box" systems (e.g., LIME or SHAP).

The research paper [12] presents LATEC, a large-scale benchmark. This benchmark tries to reconcile inconsistencies in the evaluation of Explainable AI (XAI) approaches, notably focused on saliency maps in computer vision. The "gist" of the paper can be summarized through its core contributions and findings:

The problem

Inconsistent Evaluations: A small number of XAI techniques, datasets, and metrics are often used in current research, which produces inconsistent findings.

Selection Bias: Many studies only use one or two metrics to assess a criterion, which might result in XAI method rankings that are unreliable and "overfitting" to a certain viewpoint.

The Solution: The LATEC Benchmark Massive Scale: Evaluates 17 XAI methods using 20 distinct metrics across different model architectures and input modalities

Diverse Criteria: Methods are evaluated based on three pillars: Faithfulness (does the explanation reflect the model's logic?), Robustness (is the explanation stable?), and Complexity (is it human-understandable?).

Robust Ranking: The authors propose a "rank-then-aggregate" scheme to ensure results are not skewed by outliers or different metric scales.

Important Takeaways & Insights from the article are

Top Performer: Expected Gradients (EG) identified as the most reliable strategy for both fidelity and robustness across numerous conditions. Metric Disagreement: The study reveals a high likelihood of conflicting measurements. Metrics with similar mathematical designs tend to rank ways similarly, which can disguise the true performance of a XAI technique if variety is insufficient in the evaluation set. Attention vs. Attribution: On Transformer architectures, the authors recommend using Relevance-filtered attention (LA) over standard LRP or raw attention, as it performs better in terms of faithfulness and robustness. Method Diversity: To obtain a more varied and precise knowledge of model behavior, the authors encourage practitioners to combine several approaches.

Research Design

Common sources of assessment ambiguity – there are a variety of factors that are responsible for ambiguous assessments especially in academic assessment. Sometimes the ambiguity is due to lack of clarity in the assessment guidelines and the criteria. Some other time it can be due to the subjectivity in the assessment or due to some other biases. Few common causes of ambiguity in the assessment are

Rubric vagueness: raters' interpretations of qualitative descriptors, such as "adequate analysis," vary significantly from one person to the other.

Rater subjectivity and severity: systematic bias and inconsistency are introduced when teachers consistently score higher or lower than their colleagues.

Peer-review variance: significant disagreement adds more uncertainty when using numerous peer evaluations, that is when content is evaluated by different people in the same workgroup and if the assessment varies significantly, it brings about more doubt and lack of trust and consequently the assessment becomes controversial and nit acceptable.

Opaque automated scoring: Although black-box NLP models used for essays or brief responses may be accurate, they do not provide explanations, which might lead to mistrust and inexplicable grade changes.

This study utilizes a design science research methodology combined with an experimental evaluation framework to build and validate an Explainable Artificial Intelligence (XAI)–based academic evaluation model. The methodological purpose is two fold: (1) to build an interpretable, data-driven assessment framework capable of minimizing ambiguity in academic evaluation, and (2) to empirically test the efficiency of XAI mechanisms in identifying and mitigating causes of assessment ambiguity. In order to ensure both technological rigor and pedagogical relevance, the suggested methodology incorporates explainability methodologies, machine learning, and educational measurement concepts.

To enable controlled experimentation and reproducibility, a synthetic academic assessment dataset was built, consisting of N=200 student instances. Synthetic data were selected to allow explicit modification of ambiguity-inducing variables, such as rater severity and peer-review variation, which are difficult to separate in real-world datasets. Each student record comprises structured and semi-structured attributes representing academic performance, assessment criteria, and evaluator behavior.

Feature Space Definition

Let each student instance be represented as a feature vector:

$$x_i \in R^d$$

The feature space is decomposed into three semantically meaningful subsets:

$$x_i = \left[x_i^{(A)}, x_i^{(R)}, x_i^{(T)} \right]$$

where:

x(A) denotes academic history features (e.g., prior grades, attendance rate),

x(R) denotes rubric-based evaluation features (e.g., content quality, argumentation),

x(T) denotes text-derived indicators (e.g., essay cohesion scores).

Additionally, evaluator-related attributes such as teacher severity indicators and peer-review variance were included to model procedural ambiguity.

Model Selection

An interpretable decision tree classifier was selected as the primary evaluation model. The choice of decision trees was motivated by their inherent transparency, rule-based structure, and suitability for high-stakes decision-making environments such as education.

The model learns a function:

$$F : R^d \rightarrow R$$

Such that

$$\hat{y}_i = f(x_i)$$

Tree depth was constrained to preserve interpretability and prevent overfitting.

Local Explanation Formulation

For each student instance xi, the prediction is decomposed into feature-level contributions:

$$f(xi) = \sum_{j=1}^{d} \phi j(xi)$$

Where,

ϕj(xi) denotes the contribution of feature j to the predicted outcome.

This decomposition provides instance-level transparency.

Global Feature Importance

Global feature importance is defined as:

$$I_j = \mathbf{E}_x \sim D[|\phi_j(x)|]$$

This measure identifies dominant grading factors across the dataset.

Ambiguity Quantification

Academic ambiguity is defined as:

$$Ai = \alpha\sigma_r + \beta\delta_{hm} + {}_{\gamma}S_f$$

where:

σr: inter-rater variance,

δhm : human–model disagreement,

Sf : feature sensitivity,

α,β,γ: weighting coefficients.

An instance is flagged as ambiguous if:

$$A_i > \tau$$

mapping student features to predicted academic outcomes.

Tree depth was constrained to prevent overfitting and to preserve interpretability.

The dataset was partitioned into training and testing subsets using a standard 75:25 split. Model performance was evaluated using classification accuracy and F1-score, while interpretability was assessed through rule extraction and explanation fidelity.

Results and discussions

Strong prediction performance and great interpretability were shown by the suggested XAI-based evaluation model. On the test dataset, the model's overall classification accuracy using a constrained decision-tree classifier was about 88%. This performance suggests that interpretable models need not need opaque, very complex structures to attain competitive prediction capacity. Importantly, performance stability was seen across various random divisions of the dataset, demonstrating that the model's behavior was not driven by spurious correlations or overfitting. These results illustrate the possibility of deploying interpretable models in high-stakes educational environments, where transparency is often favored over minor advances in predicting accuracy.

Global Explanation Analysis and Dominant Assessment Factors

The most significant factors influencing grading outcomes were essay cohesiveness, rubric content quality, and prior academic success, according to a global explanation analysis employing permutation-based feature importance. Attendance revealed a minor but persistent effect, while peer-review variation and teacher severity indications predominantly influenced ambiguity rather than ultimate outcomes. These results are consistent with educational

measurement theory, which highlights the importance of coherent reasoning and content mastery in academic evaluation. Simultaneously, the findings show how XAI may objectively verify—or refute—implicit presumptions ingrained in grading procedures and rubric design.Notably, characteristics linked to evaluator behavior played a substantial effect in explanation variability, highlighting their relevance in assessment ambiguity even though they were not dominating predictors of results.

Local Explanations and Instance-Level Transparency

Local explanation analysis gave fine-grained insights into individual grading choices. Feature-level contributions for each student instance made it evident how evaluator-related criteria and scholarly evidence came together to achieve the ultimate expected result.In other instances, local explanations showed that evaluator severity or peer-review disagreement was the main reason why students with similar academic profiles earned different ratings. Such findings are particularly relevant, as they expose hidden procedural errors that are difficult to identify using typical evaluation audits.From a pedagogical standpoint, these instance-level explanations encourage meaningful communication between teachers and students, making it possible to defend grades and provide helpful criticism.

Quantification and Detection of Academic Ambiguity

The proposed ambiguity metric successfully identified assessment instances characterized by high inter-rater variance, substantial human–model disagreement, and elevated feature sensitivity. Approximately 15–20% of assessment instances exceeded the predefined ambiguity threshold and were flagged for further review.

These flagged cases exhibited systematic patterns, including disproportionate influence of evaluator severity indicators and unstable rubric-driven feature contributions. This result confirms that ambiguity is not uniformly distributed across assessments but is concentrated in specific items, rubrics, or evaluator behaviors.

The ability to operationalize ambiguity as a measurable construct represents a key contribution of this work, enabling institutions to move from anecdotal concerns toward data-driven quality assurance.

Impact of XAI on Ambiguity Mitigation

The integration of XAI mechanisms considerably boosted the interpretability and diagnostic capabilities of the evaluation process. By openly disclosing feature contributions and grading logic, the approach reduced epistemic opacity and facilitated targeted intervention tactics such as rubric revision, double marking, and rater recalibration. Compared to existing automated grading procedures, which often offer limited post-hoc reasoning, the suggested approach enables preemptive detection of problematic assessment configurations. Fairness and institutional accountability will be significantly impacted by this change in assessment governance from reactive to preventive.

Implications for Educational Practice

The results of this study show that rather than serving as a decision-replacement system, XAI can serve as a decision-support tool. The framework is in line with best practices in the responsible use of educational technology by maintaining human authority while enhancing it with clear explanations and diagnostic signals. Additionally, the model offers a scalable method for incorporating explainability into learning management systems, allowing for ongoing assessment quality monitoring without placing an undue burden on teachers' cognitive capacities.

Limitations

Despite its contributions, the study has several limitations. The use of synthetic data, while enabling controlled experimentation, may not capture the full complexity of real-world assessment environments. Furthermore, only one interpretable model class was examined in the evaluation; results may differ when using different model architectures or explanation strategies. Additionally outside the purview of this study, user-centered assessment of explanation usability and educator trust merits more research.

Conclusion

In conclusion, the findings show that the suggested XAI-based assessment paradigm provides practical insights to reduce academic ambiguity while striking a compromise between interpretability and predictive efficacy. The framework helps create more equitable, reliable, and consistent academic evaluation systems by making assessment procedures transparent and diagnostically rich.

REFERENCES

[1] Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, “Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda,” in Proceedings of the 2018 CHI Conference on Human Factors in

Computing Systems, in CHI '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–18. doi: 10.1145/3173574.3174156.

[2] B. Shneiderman, "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy," Int J Hum Comput Interact, vol. 36, no. 6, pp. 495–504, 2020, doi: 10.1080/10447318.2020.1741118.

[3] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," arXiv: Machine Learning, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:11319376

[4] Naveed, S.; Stevens, G.; Kern, D.-R. An Overview of Empirical Evaluation of Explainable AI (Xai): A Comprehensive Guideline to User-Centered Evaluation in Xai. Preprints 2024, 2024100098. https://doi.org/10.20944/preprints202410.0098.v1

[5] T. Herrmann and S. Pfeiffer, "Keeping the organization in the loop: a socio-technical extension of human centered artificial intelligence," AI Soc, vol. 38, no. 4, pp. 1523–1542, 2023, doi: 10.1007/s00146-022-01391-5.

[6] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," Information Fusion, vol. 76, pp. 89–106, 2021, doi: https://doi.org/10.1016/j.inffus.2021.05.009.

[7] Goodman B. and Flaxman S., European Union regulations on algorithmic decision-making and a "right to explanation", AI Magazine. (2017) 38, no. 3, 50–57, https://doi.org/10.1609/aimag.v38i3.2741.

[8] Danks D. and London A. J., Regulating autonomous systems: beyond standards, IEEE Intelligent Systems. (2017) 32, no. 1, 88–91, https://doi.org/10.1109/MIS.2017.1, 2-s2.0-85013322063.

[9] Guidotti R., Monreale A., Ruggieri S., Turini F., Giannotti F., and Pedreschi D., A survey of methods for explaining black box models, ACM Computing Surveys. (2019) 51, no. 5, 1–42, https://doi.org/10.1145/3236009, 2-s2.0-85052502285.

[10] Classifying XAI Methods to Resolve Conceptual Ambiguity, author - Lynda Dib and Laurence Capus, Technologies, year 2025, url https://api.semanticscholar.org/CorpusID:281122150

[11] Dib, L.; Capus, L. Classifying XAI Methods to Resolve Conceptual Ambiguity. Technologies 2025, 13, 390. https://doi.org/10.3390/technologies13090390

[12] Ai, L. (2024, September 25). Navigating the Maze of Explainable AI: A Systematic approach to evaluating methods and metrics Quick review. Liner. https://liner.com/review/navigating-the-maze-of-explainable-ai-a-systematic-approach-to