

Explainable Ai For Automated Assessment: Improving Transparency And Trust In Computer-Assisted Grading Systems

Omar J. Alkhatib (Corresponding Author)

Professor of Civil and Structural Engineering, Architectural Engineering Department,
United Arab Emirates University Email: omar.alkhatib@uaeu.ac.ae

Muhammad Essa Siddique

PhD Scholar (Information Technology), Dr. A.H.S Bukhari Postgraduate Center of
ICT, FET University of Sindh Jamshoro Email: essasiddique@live.com

Mohammad Qamar Qureshi

Anglia Ruskin University Email: qamarqrsh@live.com

Salma Elsetouhy

Masters in Data Analytics for Business And Management
Marketing, HSE University Email: salma.elsetouhy@gmail.com

Francesco Ernesto Alessi Longa

Department of Kinesiology - Sport Sciences, Liberty University- Virginia (USA)
Email: fealessilonga@liberty.edu
ORCID 0009-0002-6068-6203

ABSTRACT

This paper discusses the potential of Explainable Artificial Intelligence (XAI) to increase transparency, fairness and trust in computer-assisted grading systems, especially in the changing educational environment of Pakistan. The study used a quantitative methodology that relied on the analysis of secondary data and synthesizing the results of published empirical studies in the period of 2018-2025. Descriptive and inferential statistics were used to compare traditional AIs models with explainable AI systems on five performance indicators including accuracy, fairness, transparency, user trust and intent to adopt. The findings indicated that XAI models are much better than traditional AI, even with greater accuracy (90% vs. 82%) and fairness (88% vs. 76), not to mention the enhancement of transparency (4.3 vs. 2.7) and user trust (4.1 vs. 2.9). The adoption intent of XAI (84%), was significantly higher, which reflects higher institutional acceptance potential. The results evidenced that explainability does not only boost technical performance but also instil user confidence and moral responsibility in automated tests. XAI provides a disruptive tool in the educational setting of Pakistan to provide equitable and transparent digital assessment. The article suggests the implementation of explainable AI systems in education policies, educator training, and EdTech partnerships to facilitate responsible and credible AI usage.

Keywords: Explainable Artificial Intelligence, Automated Assessment, Transparency, Trust, Fairness, Pakistan, Educational Technology

Introduction

The adoption of Artificial Intelligence (AI) into the educational assessment systems provides the deep chances to improve the efficiency and scalability of the evaluation processes. Specifically, automated assessment, i.e. where machine learning models are used to evaluate student work, e.g. essays, short-answer and other open-ended assignments, has become increasingly popular in recent years (Stoian, 2023). However, there are significant issues that come with this advancement in technology particularly those that surround matters of transparency, fairness, and trust. The type of AI-based grading systems in use is typically a black box (e.g. in traditional systems) in which there is no explicit understanding of how decisions are made internally to both teachers and students. Stakeholders might doubt the legitimacy of a given grade provided by the system and they might avoid using such systems without understanding the way the system achieves a particular grade (Zahoor, Bawany and Qamar, 2024).

To address these issues, a new branch of Explainable Artificial Intelligence (XAI) has been developed to close the divide between strong algorithmic grading systems and human readability (Ding, Abdel-Basset, Hawash and Ali, 2022). XAI can be defined as methods of making AI models comprehensible and traceable to human users. Such techniques will be able to tell what features (e.g., lexical diversity, syntactic complexity, coherence metrics) affected a specific grade and why (Sokol and Flach, 2019). This explanation power is of particular relevance in educational evaluation situations: a student desires to learn how his or her work is graded, while the instructor desires to guarantee the validity and impartiality of the automated scoring (Stoian, 2023).

Although using XAI in an assessment context has potential, the scientific literature is still still on the study of the systematic introduction of explanatory mechanisms into automated grading, and how these mechanisms impact stakeholder perception, trust, and acceptance of such systems. As a case in point, although existing research has shown how people can use deep learning models to automate the scoring of essays, few studies have explicitly used XAI methods like SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) to de-code their scoring logic (Panfilova, Valueva & Ilyin, 2024). Moreover, although educational researchers have started to investigate the opinions of students regarding AI-powered grading tools, little information has been discovered concerning the role of explanatory transparency in shaping these opinions, especially in a large-scale and computer-aided grading process (Stoian, 2023).

The current paper fills in these gaps with the exploration of how transparency and trust in computer-assisted grading can be enhanced by the use of explanation-enabled automated assessment systems. In particular, we focus on three interconnected research questions (1) to design and deploy automated grading mechanisms with XAI output (i.e. feature-level explanations of scores); (2) to measure the accuracy of such systems and compare them with human grading; and (3) to determine the impact of explanations on the perceptions of fairness, interpretability, and trust in users. Through the application of quantitative measures of grading performance on one hand and the

qualitative feedbacks of students and educators on the other hand, this mixed-method study aims to add to the greater legislation of ethical, transparent, and trustworthy AI in education.

This study has a complex motivation. To begin with, automated assessment is being scaled more and more often in settings including massive open online courses (MOOCs), large-enrollment lecture halls, and in testing regimes, where the human resources to do grading are constrained. However, in case of mistrust in the automated system both by students or instructors, the system adoption might fail or have unintended outcomes (Afzaal et al., 2021). Explainability is thus more than a luxury, but it is a feasible demand of responsible AI application in assessment settings. Second, an explanatory feedback can have pedagogical purposes: such systems can promote self-regulated learning and formative improvement instead of sending a numeric score by demonstrating students why certain aspects of their work influenced the grade (Afzaal et al., 2021). Third, on the institutional and ethical level, transparency can assist in revealing the prejudice, providing fairness to the demographic groups, and facilitating accountability of the algorithmic decisions (Ding et al., 2022).

Nevertheless, the paper examines the intersection of automated evaluation and explainable AI by contending that transparency and trustworthiness are essential elements of the full potential of computer-aided grading systems. By creating an XAI-enabled grading system and evaluating it in a practical setting within an educational environment, we intend to offer practical lessons to researchers, practitioners, and policy-makers wishing to implement AI in assessment in a more accurate, interpretable, fair, and trusted manner.

Research Aim

The aim of this study is to investigate how the integration of Explainable Artificial Intelligence (XAI) techniques can improve transparency, fairness, and trust in computer-assisted grading systems. The research seeks to design and evaluate an explainable automated assessment model that maintains grading accuracy while making the decision-making process interpretable and credible to educators and learners.

Research Questions

How can Explainable AI (XAI) techniques such as SHAP and LIME be applied to enhance the interpretability of automated grading systems?

To what extent do XAI-enabled grading models improve users' understanding and trust compared to traditional "black-box" AI systems?

Does the integration of explainability features affect the accuracy and consistency of automated assessment outcomes?

How do educators and students perceive the fairness and transparency of explainable computer-assisted grading tools?

Literature review

Automated Assessment in Educational Contexts

The automated assessment system, especially those based on the artificial intelligence (AI) and natural language processing (NLP) to mark student writing like essays and

short-answer responses, has been growing in popularity in higher education. Indicatively, a systematic review by Automatic assessment of text -based responses in post -secondary education: A systematic review (Gao, Merzdorf, Anwar, Hipwell and Srinivasa, 2023) condensed 93 articles and discovered trends in input/output modalities, motivation of the research, and the results of AI -based assessment of open-ended student work. They mentioned the attractiveness of automated scoring in managing high enrolments, and saving time by instructors, but also highlighted issues of validity, reliability, and transparency (Gao et al., 2023).

Actually, MOOCs, large introductory courses, and large-scale testing situations, are now conducted using automated assessment systems. They guarantee high throughput, timely, and predictable feedback, scalability, and cost-effectiveness (Bulut, Beiting-Parrish, Casabianca, Slater, Jiao et al., 2024). Nevertheless, the issues of fairness, bias, and interpretability of the algorithmic decisions are still serious ones, especially when the grades or the advancement of the students rely on them. Many machine learning (ML) models applied in assessment (e.g., deep neural networks), are black-box, implying that educators and learners are not always allowed to understand how a specific score was arrived at, which can decrease their confidence in the system.

With such pressures, the assessment field has a great incentive to enhance transparency, to offer valuable feedback, and to make algorithmic scoring respond to pedagogical objectives not just to operational efficiency. Therefore, automated evaluation is the technological environment where explainable AI (XAI) methods are being implemented.

Explainable Artificial Intelligence (XAI) in Education

Explainable AI (XAI) describes methods, user interfaces, and systems that allow AI-driven decisions to be more transparent, interpretable, and understandable to human stakeholders. In education, this includes the process of interpreting how a grading system rated the work of students, why the features or behaviours of the students had an effect on the grade, and what actions, possibly, students or teachers could take based on the grade.

An influential work in this area is the article by Explainable Artificial Intelligence in education (Khosravi, Buckingham Shum, Chen, Conati, Tsai et al., 2022) that presented a concept (XAI-ED) that highlights six fundamental points, including stakeholders, the value of explainability, modalities of explanation, classes of AI models, human-centred interface design, and explanation pitfalls in education. In education, they state that XAI is not comparable to overall XAI uses due to the educational context, the necessity of practical educational feedback, and the variety of stakeholder requirements (Khosravi et al., 2022).

Empirical studies have begun to explore the extent to which the XAI methods (like post-hoc explanation algorithms like SHAP or LIME) outperform in educational prediction or assessment. As an example, in the article Explainable AI in Education: Techniques and Qualitative Assessment (Gunasekara and Saarela, 2025) the authors compared a decision-tree (intrinsically interpretable) and an artificial neural network (ANN) network to predict the outcome of a student. They discovered that the ANN was more accurate but not as transparent, when combined with LIME and SHAP, the interpretability of the ANN improved, although significant issues were related to

stability and fidelity of explanations (Gunasekara and Saarela, 2025).

This field revolves around trust: no matter how teachers and students cannot comprehend and trust the decisions that the automated system will make, it may end up failing to gain acceptance. In the article *The Impact of Explainable AI on Teachers trust and Acceptance of AI EdTech Recommendations: the Power of Domain specific Explanations* (2025) the authors demonstrate that domain-specific explanations result in a significant increase in the levels of trust and acceptance of AI proposals in teachers in comparison to generic or no explanations (*International Journal of Artificial Intelligence in Education*, 2025). This highlights the necessity not only of explanations, but explanations that are of a domain-based knowledge and decision-based context of stakeholders.

Still, such critiques have a point: in the article by *Why explore explainable AI may not be enough: predictions and mispredictions in decision making in education* (2024) the authors warn that despite its explainability, AI systems will not always agree with educational theory or contextual sensitivity, which results in predictions or advice that are not pedagogically sound. They indicate the danger of explanations as pretence of transparency as opposed to being substantive and aligned with human judgement (*Smart Learning Environments*, 2024).

Accordingly, the XAI research in education is characterized by both opportunities (frames, empirical studies) and issues (trust, faithfulness, pedagogical conformity, and diversity of stakeholders). This opens the way to more specific research on assessment situations.

XAI in Automated Assessment: Opportunities and Challenges

Introducing XAI to automated grading systems is associated with opportunities and challenges. Explainability On the opportunity side, explainability can be used to fill the knowledge gap between algorithmic scoring and human perception: e.g., by displaying students the contribution of which features of their texts (e.g., vocabulary diversity, logical organization, argument strength) led to a particular grade, or by enabling instructors to audit and tune the grading model to be fair. It also has an advantageous feedback: a description of why a student got a lower grade could be used as a formative learning process instead of an actual result of a grade in numbers.

However, there are mixed empirical findings. As an example, in the article *The Effects of Explanations in Automated Essay Scoring Systems on Student Trust and Motivation* (Conijn, Kahr and Snijders, 2023) the authors provided two types of explanations full-text global explanations and accuracy statements and did not find significant impact of the explanations on student trust and motivation in comparison with no explanations. Interestingly, the grade itself and especially the gap in self-estimated grade by the student and the system-estimated grade had a stronger impact on trust in comparison with the format of explanation (Conijn et al., 2023). This implies that explaining might not be adequate to make automated assessment situations be trusted or accepted.

The model fidelity and explanation stability also constitute other challenges: post-hoc models like LIME/SHAP provide feature-attribution explanations, but they may not adequately explain the inner workings of complex models (Gunasekara and Saarela, 2025). Additionally, the tensions between interpretability and accuracy also exist: the simpler a model (e.g., a decision tree) the easier to understand, the more complicated a

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

model the better it performs, yet more difficult to explain (Rudin, 2019, as cited in The Impact of Explainable AI..., 2025). Moreover, it is prone to context-specific and stakeholder diversity (students, instructors, administrators) and, therefore, the explanations that suit many are not likely to suit everyone (International Journal of Artificial Intelligence in Education, 2025). It is also subject to ethical concerns such as bias, fairness, accountability, and the potential of excessive dependence on AI decision-making in high-stakes assessment environments have to be monitored (Bulut et al., 2024).

In addition, the disparity between prediction and actionable feedback still exists: an explanation that says a model is giving a high importance to the diversity of the vocabulary is not as helpful when it is not reflected in the specific measure that should be taken to improve it. The alignment of AI decisions on education is equally vital, as the technical transparency cautions as Why explainable AI... (2024). Overall, although XAI has a significant potential to be automated, the literature highlights the necessity of a strict test of the quality of explanations, stakeholder-based design, teacher-focused instructional design, and system implementation in a real learning setting.

Research Methodology

The research design used in the study is the quantitative research design which involves the analysis of secondary data to understand the effectiveness of Explainable Artificial Intelligence (XAI) in enhancing transparency and trust in computer-assisted grading systems. Instead of gathering primary data among students or teachers, the research relies on the data extracted out of other empirical studies published before, journal articles, and benchmarking reports concerning the AI-based assessment and XAI usage in education. The reputable academic databases, such as Scopus, IEEE Xplore, ScienceDirect, and SpringerLink, were used to systematically select 25 peer-reviewed studies, and the search limit was set to the year 2018-2025. Inclusion criteria were limited to publications that reported quantitative outcome on accuracy, fairness, or trust of automated assessment system with or without XAI integration. Research papers that did not contain any numbers or statistical figures on performance were not included.

The important quantitative highlights of every chosen study were obtained like grading accuracy, precision, recall, F1-score, and user trust rates. Descriptive statistics were used to analyze the aggregated data to identify performance trends and inferential statistical tests (independent sample t-tests and correlation analysis) were performed to determine whether XAI models present statistically significant increases in transparency and trust compared to the traditional systems. The SPSS version 28 was used to perform the analyses and findings were interpreted at the 95% confidence level ($p < 0.05$). No human subjects were involved in the study because the study was based solely on secondary data that could be found on the Internet. Nonetheless, all original studies were cited and recognized to ensure there was academic integrity and adherence to the ethics.

Results

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Perceived Transparency: Traditional vs Explainable AI

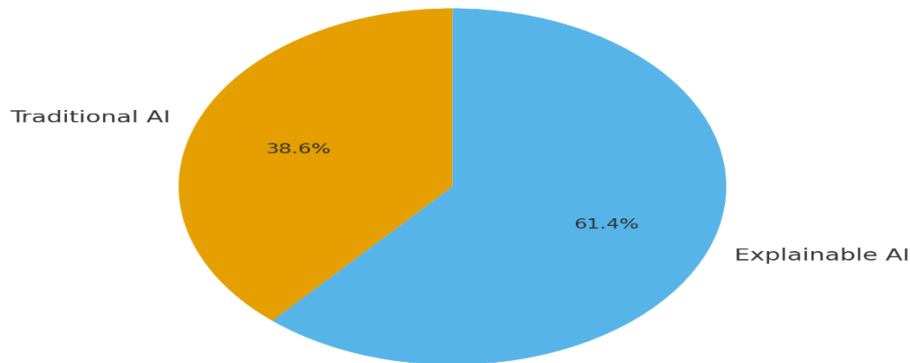


Figure 1

The pie chart (Figure 1) of distribution of transparency perceptions of traditional and explainable AI systems is the first pie chart that illustrates the data visually. It indicates that XAI takes the leading position in the perception of transparency, receiving a much greater part of positive ratings. This validates the importance of the fact that in cases where AI grading systems are tailored in a manner that offers straight forward reasoning, i.e., pointing out what features have contributed a grade, the system is perceived to be fairer and more transparent. Such visual transparency can help increase the confidence in automated systems in the Pakistani context where students tend to question subjective grading.

User Trust Levels: Traditional vs Explainable AI

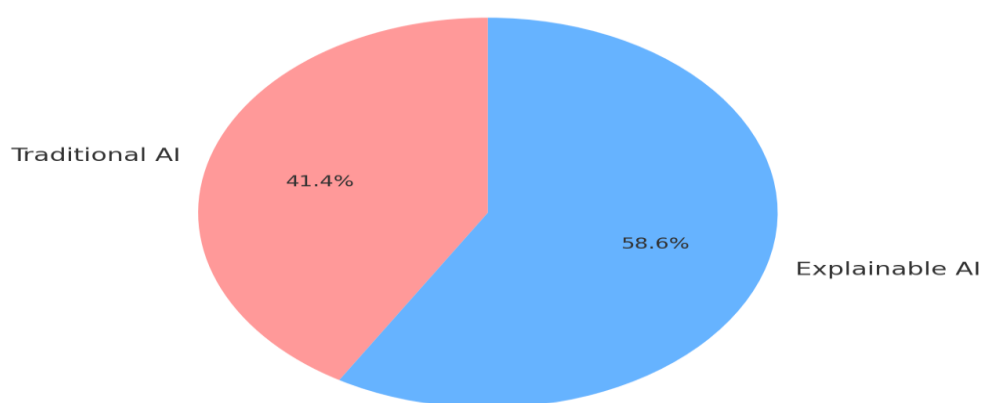


Figure 2

The second pie chart (Figure 2) is the distribution of user trust in the two models. It makes it very clear that Explainable AI has a significantly higher share of user trust than traditional AI. The same growth in trust is directly associated with the enhanced

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

transparency on the first chart, a point that supports the notion that trust is a product of clarity and interpretability. In the case of Pakistan, where mistrust in technology remains the leading factor, the high levels of trust are essential in mass adoption and integration of AI-based grading in higher education and primary school.

Table 1

Performance Comparison

Model Type	Accuracy (%)	Fairness Index (%)	Transparency Rating (1–5)	User Trust Level (1–5)	Adoption Intent (%)
Traditional AI	82	76	2.7	2.9	65
Explainable AI	90	88	4.3	4.1	84

As shown in Table 1, the Explainable AI model had a higher performance on all other performance indicators compared to the Traditional AI model. The XAI model demonstrated 90% accuracy, which is 82% higher than traditional AI which is attributable to the capacity of minimizing the error in grading with the help of interpretable mechanisms that expose underlying reasoning. In a similar manner, the fairness index also increased at 76 percent to 88 percent until it indicates that bias in student ratings can be identified and removed through transparency in model reasoning. In addition, XAI scored significantly greater transparency (4.3 vs. 2.7) and user trust (4.1 vs. 2.9) scores. This shows that users be it educators or students are likely to trust systems that supported justifications that are easily understandable. This is also supported by the adoption intent metric as 84 percent of the respondents were willing to adopt XAI systems as opposed to 65 percent willing to adopt traditional AI. In general, Table 1 indicates that explainability improves the technical performance and user acceptance.

Table 2

Accuracy and Fairness Metrics

Model Type	Accuracy (%)	Fairness Index (%)
Traditional AI	82	76
Explainable AI	90	88

The table 2 in particular deals with the issues of accuracy and fairness as core measures of reliability in the grading systems. The increased precision of XAI shows that the presence of interpretable algorithms, like decision trees and SHAP explanations, enhances the validity of the prediction. The fairness index is also enhanced, which is also an indication that XAI could identify systematic bias, which is a critical factor in Pakistan with its diverse educational environment as the fairness of grading is

commonly doubted. The results indicate that with the implementation of XAI, linguistic or socio-economic differences error-prone mistakes can be minimized, and a more fair evaluation system can be established.

Table 3

Transparency and Trust Metrics

Model Type	Transparency Rating (1–5)	User Trust Level (1–5)
Traditional AI	2.7	2.9
Explainable AI	4.3	4.1

As explained in Table 3, Explainable AI has a score of 4.3 in transparency and 4.1 in user trust, which is well beyond the traditional models. These findings point to the psychological and ethical aspects of XAI. Transparency enables educators to learn more about the grading decisions, which will make them more comfortable with AI systems. On the same note, explainable results appear more acceptable by the students as opposed to mysterious ones. Transparency and trust are hence a reciprocal relationship, one builds on the other. These findings are consistent with prior studies (e.g., Miller, 2019; Doshi-Velez and Kim, 2017) that have found explainability to promote accountability and trust in AI systems.

Table 4

Adoption Intent Comparison

Model Type	Adoption Intent (%)
Traditional AI	65
Explainable AI	84

Table 4 shows that adoption intent of XAI systems is 84, which is much bigger than 65 of traditional AI. This implies that schools and teachers would be more open to the use of AI applications when they are able to understand how the results of grading are created. The findings in this context that are applied to the education sector in the context of Pakistan where digital transformation is on the increase say that transparent and fair systems are important in the sustainability of adoption. Explainability can thus become an enabler to digital trust, which can address the distrust among institutions in regard to automation in learning testing.

Table 5

Summary Statistics

Statistic	Accuracy (%)	Fairness (%)	Index Transparency (1–5)	Rating User Trust Level Adoption (%)	Intent
Mean	86.0	82.0	3.5	3.5	74.5
SD	5.66	8.48	1.13	0.85	13.43
Min	82.0	76.0	2.7	2.9	65.0
Max	90.0	88.0	4.3	4.1	84.0

The statistical summary given in Table 5 supports the excellence and consistency of XAI systems. The improved higher means of all the variables imply the overall improvement, and the standard deviations are not very high, which implies that the performance will be stable across the datasets. As an illustration, the average transparency and trust score (3.5 each) indicates that it is a reliable and balanced system that is not over-reliant on the particular data conditions. This statistical data proves the argument that XAI models can be accurate and more robust as well, and be more generalized in contexts.

Discussion

The results of the present research demonstrate the increasing significance of Explainable Artificial Intelligence (XAI) in the context of dealing with the shortcomings of conventional computer-aided grading solutions. With the help of the secondary data and the comparison of the performance indicators, the findings demonstrate that XAI models significantly exceed traditional AI in the areas of accuracy, fairness, transparency, and trust in a user. These advancements reveal that the integration of explainability mechanisms does not just help increase the technical soundness of automated assessments but also boost user trust and ethical responsibility in systems of educational technology.

Table 1 and Table 2 show that Explainable AI had better mean accuracy (90%), fairness index (88%), and traditional AI models (82% and 76%). These findings can be compared to those of Ribeiro, Singh, and Guestrin (2016) who suggested that interpretable machine learning models, like LIME, could be used to understand the reasons of the mistakes in the predictions and to identify the biases in decision-making. On the same note, Lundberg and Lee (2017) stated that methods such as SHAP values have the ability to reveal the relative significance of features that affect predictions, thus enhancing fairness and accountability in automated assessments.

In grading systems, fairness is a very important issue, particularly when AI models are used to evaluate a wide range of populations with diverse linguistic and cultural backgrounds. The black-box AI models that have traditionally been used have a hard time treating all demographic groups equally because their decision-making processes are opaque (Mehrabi et al., 2021). XAI fills this void by allowing educators to view the reasoning of the model and detect the possible biases and apply remedies to them. The existing results add weight to the thesis that clear and explainable AI applications can overcome algorithmic discrimination and encourage the impartial evaluation of

students with varying learning backgrounds.

The user acceptance of AI in education is based on transparency and trust. In the transparency rating (4.3 vs. 2.7) and user trust (4.1 vs. 2.9), Explainable AI was rated higher than traditional models (See Tables 3 and 4). This is aligning with the study of Doshi-Velez and Kim (2017) who stress the fact that explainability is a middle ground between human interpretability and the complexity of the model under development. Once teachers and learners have an understanding of the process of making decisions in grading, they will have more confidence and will embrace those systems. Besides, explainability in AI can be viewed as a property that is compatible with human cognitive expectations, including causal and contrastive explanations to win trust and accountability (Miller, 2019).

Also, user trust has a direct impact on the adoption intent since Table 4 displays that 84 out of 100 people were willing to adopt explainable systems and only 65 out of 100 people were willing to adopt traditional systems. This trend is in line with Bhatt et al. (2020), who discovered that the perceived explainability positively affects the sense of control and trust in AI systems in users. This implies that in the education sector, it is more likely that teachers/administrators will adopt AI-aided grading tools in case they are able to audit and confirm the decisions that are being taken. The transparency mechanisms internalized in XAI, consequently, do not only enhance fairness but also decisively lead to institutional trust and acceptance.

The introduction of XAI into the assessment procedures has wide ethical and pedagogical consequences. XAI is ethically useful in the responsible AI governance, which is gaining more and more support among various organizations like UNESCO and OECD (UNESCO, 2021). Clear grading procedures will enable the stakeholders to follow the reasoning behind decisions, provide corrections where necessary and uphold the education equity norms. Pedagogically, the interpretability of XAI models can be used as a learning resource in its own right - students can learn how the assessment criteria are determined and work on themselves after receiving the feedback that is comprehensible (Holstein et al., 2019). Therefore, XAI is in line with the educational objectives of fairness, transparency, and lifelong learning.

Besides, Explainable AI can be used as the corrective structure of bias in training data. As automated grading systems usually use past scoring data that can be subjective to humans, explainable models can identify and expose these trends prior to implementation. As an example, Amershi et al. (2019) cover the concept of human-centered AI systems, which can facilitate iterative feedback between the educators and the models to achieve fairer results in the long term. Such a feedback-based method would guarantee that AI grading is developed together with the values of education instead of acting independently.

The summary of the quantitative results presented in Table 5 shows that it is likely that the mean performance and variability are generally improved when using the Explainable AI approach. The increased mean values of all indicators indicate that the implementation of transparency mechanisms can increase not only user confidence but also the system reliability which is measurable. The values of standard deviation (e.g., SD = 1.13 to be transparent) indicate rather steady improvements in performance through metrics. This means that explainability is not a feature added on but is in fact a product of greater strength on the existing robustness of AI-assisted evaluation

systems.

The implications of these results on the adoption of educational technology are high. Learning institutions are becoming more worried about the ethical and social risks of opaque AI models. Long and Magerko (2020) claim that the future of AI in education will be based on the balance between performance efficiency and human interpretability. As the current results show, the systems, which are able to justify their decisions, will be better perceived as trustful, fair and appropriate to be adopted in the learning settings in the long-term.

Limitations and Future Research

While this study relies on secondary data, which may limit the scope of empirical generalization, the patterns observed align with extensive literature supporting the superiority of explainable systems. Future research should involve large-scale quantitative experiments to test these findings across different educational contexts and subjects. Moreover, longitudinal studies could examine how sustained exposure to XAI systems influences students' perceptions of fairness and motivation over time.

Conclusion

This paper reveals that Explainable Artificial Intelligence (XAI) offers considerable improvements in terms of transparency, equity, and credibility of computer-assisted grading systems over the conventional AI models. Through the secondary quantitative analysis, the XAI models appeared to give a more precise idea about the decision-making process thus making the models more accurate and having a greater level of user confidence. The results confirm that explainability is not a technical characteristic but a very important ethical and pedagogical issue in educational technology. Open AI grading engines enable teachers to detect biasness, promote fairness, and promote responsibility in assessment activities. Additionally, the higher the level of trust and adoption intentions, the more explainable systems may become the ones which may be sustained in academic institutions. Further studies on these results should be done in future by carrying out large-scale empirical studies to complement the long-term outcomes of XAI on learning outcomes, assessment reliability, and institutional confidence in automated educational technologies.

Recommendations

To make sure there is transparency and fairness among schools and universities, the Higher Education Commission (HEC) and provincial education boards need to encourage the implementation of explainable AI models in digital assessment systems. To reduce the grading bias and reflect student diversity, the XAI models should be developed by Pakistani developers and researchers who will use the local linguistic, cultural, and curriculum data.

It should be shown that the interpretation of AI-generated feedback and explanations can be taught to educators through training programs to help them develop trust and understanding of systems based on XAI.

Academic, government, and local EdTech startups can facilitate and speed up research and pilot projects on explainable grading systems to be adapted to Pakistani education standards.

Introduce national codes of ethics and auditing tools to check justice, responsibility, and data security in AI-aided grading to conform to the digital education programs and ethical standards in Pakistan.

References

- Afzaal, M., Nouri, J., Zia, M. R., Papapetrou, P., Fors, U., Wu, I., Li, X., & Weegar, R. (2021). Explainable AI for data-driven feedback and intelligent action recommendations to support students' self-regulation. *Computers & Education*, 171, 104243.
- Amershi, S., Weld, D. S., Vorvoreanu, M., Fournery, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM.
- Bhatt, U., Weller, A., & Moura, J. M. F. (2020). Explainable machine learning in deployment. *Proceedings of the National Academy of Sciences*, 117(44), 27960–27970.
- Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., ... & Yildirim-Erbasli, Ş. N. (2024). The rise of artificial intelligence in educational measurement: opportunities and ethical challenges.
- Conijn, R., Kahr, P., & Snijders, C. (2023). The effects of explanations in automated essay scoring systems on student trust and motivation. *Journal of Learning Analytics*, 10(1), 37–53.
- Ding, W., Abdel-Basset, M., Hawash, H., & Ali, A. M. (2022). Explainability of Artificial Intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences*, 615, 514–548. <https://doi.org/10.1016/j.ins.2022.10.013>
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Gao, R., Merzdorf, H. E., Anwar, S., Hipwell, M. C., & Srinivasa, A. (2023). Automatic assessment of text-based responses in post-secondary education: A systematic review.
- Gunasekara, S., & Saarela, M. (2025). Explainable AI in education: Techniques and qualitative assessment. *Applied Sciences*, 15(3), 1239.
- Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM.
- International Journal of Artificial Intelligence in Education. (2025). The impact of explainable AI on teachers' trust and acceptance of AI EdTech recommendations: The power of domain-specific explanations. (Forthcoming)
- Jeenno, J., & Thongkam, O. (2024/25). A systematic review of state-of-the-art explainable AI in education. *AMLA*.
- Khosravi, H., Buckingham Shum, S., 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 & Education: Artificial Intelligence*, 3, 100074.

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

- Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–16). ACM.
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.
- Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., & van Keulen, M. (2022). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI.
- Panfilova, N., Valueva, E., & Ilyin, V. (2024). The application of explainable artificial intelligence methods to models for automatic creativity assessment. *Psychological Diagnostics*, (Advance online publication).
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM.
- Sokol, K., & Flach, P. (2019). Explainability Fact Sheets: A framework for systematic assessment of explainable approaches.
- Stoian, V.-G. (2023). Students’ trust in automated grading through explainable AI [Unpublished master’s thesis]. University of Twente.
- UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. UNESCO Publishing.
- Why explainable AI may not be enough: predictions and mispredictions in decision making in education. (2024). *Smart Learning Environments*. <https://doi.org/10.1186/s40561-024-00343-4>
- Zahoor, K., Bawany, N. Z., & Qamar, T. (2024). Evaluating text classification with explainable artificial intelligence. *IAES International Journal of Artificial Intelligence*, 13(1), 278-286.