Explainable Artificial Intelligence in Automated Essay Scoring: A Systematic Review of Transparency In AI-Based Educational Assessment

Kazi Siam Al Mobin[1]

1. Department of Information Technology, Washington University of Science and Technology, United States

ABSTRACT

Artificial intelligence (AI) is also an increasing trend in educational assessment due to the fact that it has enabled automated assessment systems to perform large-scale analysis of written responses. The Automated Essay Scoring (AES) systems involve the application of machine learning algorithms and natural language processing to analyze the textual features such as grammar, vocabulary, coherence and semantic relevance to generate scores that are similar to those that would be rated by a human. These systems have acquired mass applications in mass assessment of learning due to its efficiency, scale ability, and the ability to reduce the amount of time taken to mark. However, this application of AI-based methods of assessment has raised some grave concerns regarding the transparency, interpretability, and fairness of algorithmic decision-making. The existing AES models are mostly founded on detailed machine learning and deep learning models which act like a black box and an individual will therefore not fully understand how the scores are generated. The lack of transparency may reduce the trust in the automated grading systems and make the accountability in the education assessment to be called into question. To manage these problems, the concept of Explainable Artificial Intelligence (XAI) has become a valuable field of research to enhance the interpretability and transparency of machine learning models. Shapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and attention-based visualization have become popular techniques of XAI and are being incorporated into AES systems to offer insights as to how algorithms make judgments based on textual features and how they assign scores. Explainability is especially relevant in the field of education since the results of assessments determine the learning of students, their academic advancement, and the work of educational institutions. Although the variety of studies on explainability in AI-based educational tools is on the rise, the research on explainable AES continues to be divided in various fields such as educational technology, natural language processing, and learning analytics. Thus, an in-depth analysis of available information is required to comprehend the situation in the field of research and to figure out the directions of the further development. This paper is a systematic review of the literature about explainable artificial intelligence in automated essay scoring in terms of AI models applied to assess essays, explainability methods with the aim to increase transparency, and the consequences of the methods identify the effects of the methods on educational assessment.

Keywords: Explainable Artificial Intelligence, Automated Essay Scoring, Educational Assessment, Natural Language Processing, Algorithmic Transparency

INTRODUCTION

The introduction of artificial intelligence (AI) in education has completely altered how learning and assessment is done. Automated Essay Scoring (AES) is one of the most well-known examples of AI use in educational assessment, which presupposes the use of computational models to mark written answers and give them a score that is similar to that given by human raters. The AES systems apply natural language processing (NLP) and machine learning analysis to linguistic and semantic content of the student essays, such as grammar, vocabulary complexity, coherence, and argument structure [1]. These technologies have been prompted by the desire to enhance efficiency and scalability in large scale educational testing environments where human grading can be time consuming, costly, as well as prone to human variation. The automated essay scoring systems have a history of several decades. The initial versions of the AES used mostly the methods of statistics and manual linguistic characteristics to estimate the quality of the writing. Project Essay Grade (PEG) was among the oldest and most powerful systems employing the surface-level textual characteristics, like the essay length and the number of words per frequency, to approximate the quality of writing [2].

Subsequent systems like e-rater produced by the Educational Testing Service (ETS) added more advanced features to the language like grammar, usage, style and factors of discourse to be able to give better prediction of scoring [3]. It was shown by these systems that the degree of agreement with human raters in automated scoring could be as high as that between human raters themselves [4]. The developments in machine learning have also increased the power of

AES systems. The current models become more dependent on deep learning models, including recurrent neural networks, convolutional neural networks, and language models based on the transformers, to identify contextual and semantic connections within text [5]. These models can also compute more complicated linguistic patterns and compare more facets of writing like argument coherence and semantic relevance. It has been demonstrated that AES models based on deep learning can have a high predictive accuracy and close correlations with human scoring standards [6].

The combination of large-scale language model and contextual embedding further boosts the performance of automated essay scoring systems to analyze essays on a more linguistic language level and a more context-aware level [7]. Although this has been achieved, the growing complexity of the AI models have brought in questions of transparency and interpretability of automated assessment systems. Most machine learning models in the AES are black-box models, i.e. the decision-making mechanisms of the models are hard to interpret or explain [8]. This has not been seen as a favorable issue in the educational settings since the results of the assessment can affect the academic pace of advancement, scholarship prospects, and school reviews of students. Lack of transparency on a score creation process between educators and learners might lead to a loss of confidence in automated assessment technologies and questions about fairness and accountability [9]. In order to overcome these issues, scholars have resorted to a new area of study of Explainable Artificial Intelligence (XAI). XAI is a set of procedures and approaches that allow making AI models more readable and understandable by offering insights into the reasons of their predictions and decisions [10].

Within the framework of the assessment in education, explainability can assist teachers to get a clearer picture on what exactly in the text or other linguistic pattern defines the scoring decisions of a model. XAI can further increase the confidence and acceptance of AI-based assessment systems in educators and students by making algorithmic reasoning more transparent. A number of explainability methods have been suggested to enhance the transparency of automated essay scoring models. In one example, Local Interpretable Model-Agnostic Explanations (LIME) can be used to explain local predictions made by a model by determining the most influential features that can influence a particular decision [11]. Likewise, SHapley Additive exPlanations (SHAP) applies concepts of cooperative game theory in an attempt at estimating the value of individual features to the prediction of a model [12]. The techniques enable researchers and educators to see the significance of various linguistic features, hence the scoring process becomes more understandable. Moreover, neural models that focus on attention have been applied like emphasizing certain words or sentences that play a major role in the scored results, which provide additional information about the reasoning of deep-learning models internally [13].

It is becoming clear that explainability in AI-powered educational systems is a significant necessity in becoming a responsible and ethical AI implementation. Clearly defined AI systems might help create more transparent decision-making, allow teachers to authenticate the algorithmic feedback, and offer students valuable feedback concerning performance in the writing [14]. Besides, explainable AI methods will enable human-AI interaction, whereby an educator may decipher model projections and incorporate them in the pedagogical decision-making routines. Although the idea of explainable AI in educational contexts has become a common topic of research, the research on explainability in automated essay scoring in particular is still disjointed. Current research is spread over various disciplines including natural language processing, educational data mining and learning analytics and it is hard to have an overview of the current events and research trends. Moreover, although a number of studies have put forward explainability methodologies in AES systems, there has been limited generalization on how this methodologies can be used in bringing transparency, trust, and equity in educational assessment [15].

In addition, despite some studies having proposed explainability methodology in AES systems, generalization of the application of the methodologies in introducing transparency, trust, and equity in educational assessment has been minimal [15]. Taking these concerns into consideration, a methodological literature analysis is needed that would

enable to consolidate the existing study and establish the significant advances in the scientific field of the explainable automated essay scoring. This study is therefore aimed at having a systematic literature review of explainable AI in automated essay grading systems in the hope of finding out what kind of AI models are implemented to carry out essay grading, the explainability mechanisms in such systems that enhance transparency and what effects such mechanisms have on educational grading. The purpose of the reviewed work is to sum up the information regarding the existing research on the subject matter to illuminate the existing tendencies in technology, the knowledge gaps, and help establish more open and reliable AI-based educational assessment systems.

METHODOLOGY

The systematic review is conducted due to the necessity to conduct a literature review on the subject of explainable artificial intelligence (XAI) methods employed in automated essay scoring (AES) systems. The review was carried out based on the recommendations of Preferred Reporting Items of the Systematic Review and Meta-Analysis (PRISMA 2020) to ensure that the process of performing the study identification, screening process, evaluation of the eligibility criteria, and inclusion provided the transparency and methodological rigor.

2.1 Review Design

The systematic literature review methodology was adopted to identify and evaluate peer-reviewed studies on the subject of the integration of explainable artificial intelligence techniques in automated essay scoring systems. To conduct the review, the studies that investigated the use of machine learning or deep learning techniques to evaluate essays and involved interpretability or explainability mechanisms to increase the transparency of AI-based educational assessment were considered.

2.2 Search Strategy

The search strategy was planned as a full search strategy to locate the pertinent literature in various academic databases. These searches were in the period in-between the years 2010 to December 2024 since this is the period when both automated essay scoring systems and explainable AI approaches have significantly evolved. The search was carried out in the following electronic databases:

- Scopus
- Web of Science
- IEEE Xplore
- ERIC
- Google Scholar

Such databases have been chosen since they all represent the study of the field of education, artificial intelligence, natural language processing, and learning analytics.

The search strategy was a combination of the keywords which were connected to automated essay scoring and explainable artificial intelligence. The use of Boolean operators and variations of keywords was done to ensure that all the relevant studies were retrieved. The key-word of the search that was applied in the databases was:

("automated essay scoring" OR "AES" OR "automated writing evaluation")
AND
("explainable artificial intelligence" OR "XAI" OR "model interpretability" OR "algorithm transparency")

AND
("education" OR "educational assessment")

Additional manual searches were conducted by examining the reference lists of relevant review articles and selected papers to identify further eligible studies.

2.3 Inclusion and Exclusion Criteria

Eligibility criteria were defined prior to the screening process to ensure consistency in study selection.

Inclusion Criteria

Studies were included if they met the following criteria:

1. The study focused on automated essay scoring or automated writing evaluation systems.
2. The research incorporated machine learning or deep learning models for essay evaluation.
3. The study implemented explainability or interpretability techniques such as SHAP, LIME, attention visualization, feature importance analysis, or other transparency methods.
4. The article was published in a peer-reviewed journal or conference proceeding.
5. The publication was written in English.
6. The study provided empirical evaluation, methodological development, or experimental analysis of AES systems.

Exclusion Criteria

Studies were excluded if they:

1. Focused on AI applications in education without addressing automated essay scoring.
2. Discussed AES systems without incorporating explainability or interpretability methods.
3. Were editorials, opinion articles, book chapters, or dissertations.
4. Did not provide sufficient methodological or experimental details.
5. Were duplicate publications across databases.

2.4 Study Selection Process

The process of selecting the study involved the four steps suggested in the PRISMA 2020 framework: identification, screening, eligibility, and inclusion. In the identification phase, 482 records were found in the sampled databases. Upon the deletion of 96 duplicate records, there were 386 distinct articles to be screened. During the title and abstract screening step, papers that were evidently unrelated to automated essay scoring or explainable AI were filtered out. This led to the exclusion of 298 articles, leaving 88 studies to fully review. The evaluation of each article during the full-text eligibility assessment was conducted against the prespecified inclusion and exclusion criteria. Articles that failed to apply explainability solutions, had no empirical testing, or were not related to educational applications of AI

were filtered out. This stage then led to the removal of 71 studies. Ultimately, 17 articles that met the eligibility criteria were included in the systematic review. These studies comprised the ultimate dataset of qualitative synthesis.

The study selection process is illustrated in the PRISMA flow diagram presented in Figure 1.

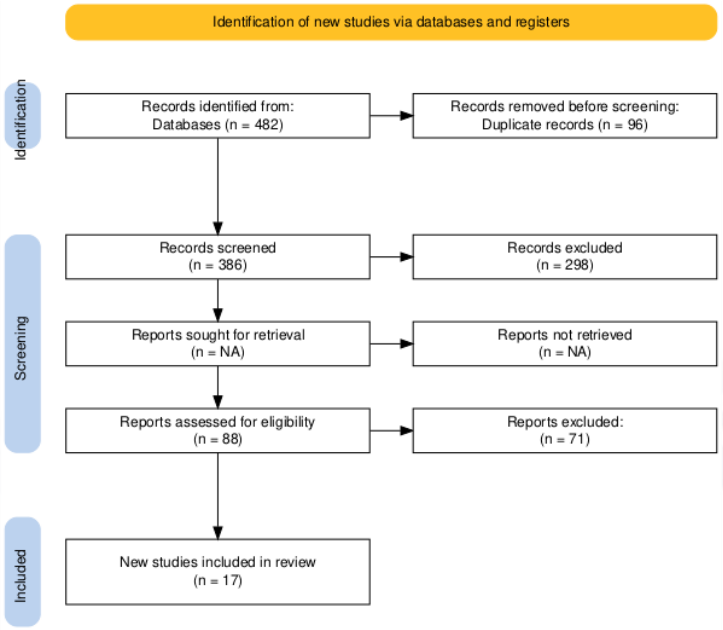

Figure 1: *Prisma Diagram*

2.5 Data Extraction

A data extraction framework was created to select the information related to every included study in a structured manner. The extraction was conducted with two key dimensions, which were the technical properties of automated essay scoring (AES) models and the methods of explainability, used in these models. The important details in each study were collected and these details included the author(s) and year of publication, the dataset by which essay evaluation was conducted, the type of artificial intelligence or machine learning model applied, the explainability method applied, and the measures used to assess the performance of the model, such as Quadratic Weighted Kappa (QWK) and correlation with human raters. Besides, the key results and contributions in regard to transparency and interpretability in AES systems have been reported. All the information which was extracted was properly structured into a summary table to allow easy comparison of the studies as well as to discover new trends in the evolution of explainable automated essay scoring systems.

2.6 Quality Assessment

All included studies underwent a quality assessment procedure to be able to guarantee the methodological rigor and reliability of the results and use modified criteria to evaluate the studies available through the Mixed Methods Appraisal Tool (MMAT). Every study was rated on four important dimensions: are the research objectives clear, is the development of the AI models and their methodology described transparently, do the evaluation metrics used to measure model performance have a good fit, and the level to which the model outputs were explained or interpreted. All of the criteria were rated on a three-point scale, 0 meaning that the criterion was not addressed, 1 meaning that the

criterion was partially addressed, and 2 meaning that the criterion was wholly addressed. In the process of selecting the studies, those that had a lower score than a stipulated quality threshold were filtered out at the eligibility phase to ensure that the overall quality of the review methodology remained uncompromised. The outcomes of the assessment showed that the majority of the studies that were included showed moderate to high methodological rigor, especially regarding the experimental design, evaluation process, and the presentation of performance measures.

2.7 Data Synthesis

Since AI models, datasets, and methods of explainability vary in terms of their methodological diversity and variation across the studies included, a qualitative approach to data analysis was employed. Data obtained were subjected to thematic analysis to determine similar trends and conceptual ideas in the literature. The synthesis was conducted in the context of four primary areas, namely the nature of AI models used to score essays automatically, the explainability methods implemented in AES systems, the implications of using explainable AI in automated assessment in education, and the limitations and gaps in research identified in existing literature. This thematic synthesis helped to gain a broad insight into the explainability of artificial intelligence being applied into automated essay scoring systems and its insights into how the technologies can be used to enhance transparency, interpretability, and confidence in AI-based educational assessment.

RESULTS

After the PRISMA screening and eligibility procedure as explained in the Methods section, 17 articles were incorporated in the ultimate qualitative synthesis. These articles were written in 2016-2024, which is why the research interest in explainable artificial intelligence in automated essay scoring systems has increased. The chosen works are an interdisciplinary work covering the areas of educational technology, natural language processing, artificial intelligence, and learning analytics. The majority of studies were devoted to the creation or testing of machine learning and deep learning algorithms to be used in automated scoring of essays with the integration of explainability systems to enhance the process of transparency in algorithms used when assessing the essay.

The research articles in included studies differed in terms of methodology, datasets, and explainability procedures. Some studies had designed neural-network-based AES models, and some had investigated transformer-based models and interpretable machine learning models. A smaller group of studies had a specific area of human-centered consideration, which is the need to enhance the trust and comprehension of educators of the automated scoring system through explainable AI outputs. All in all, the literature suggests a growing focus on the concept of transparency and interpretability in AI-based educational assessment instruments.

3.1 Characteristics of Included Studies

The included studies are randomized trials, with at least half randomized trials, and the remaining half controlled trials, with at least half controlled trials. The research has covered various data sets, both popular essay scoring datasets like the Automated Student Assessment Prize (ASAP) dataset, and institutional educational corpora. The performance of the models was evaluated in most studies via Quadratic Weighted Kappa (QWK), Pearson correlation, and root mean square error (RMSE) to compare the predictions of automated scoring and human ratings. Some studies were mainly aimed at enhancing the accuracy of scoring using advanced machine learning architectures, and others were organized on the necessity of exploring explainability methods into AES models. During the last several years, model-agnostic interpretability approaches have been more and more popular with researchers, which enables educators to visualize the textual attributes that contribute to scoring decisions. This tendency indicates the increasing acceptance of the idea of need of transparent AI systems in responsible deployment in educational assessment settings. The 17 studies that have been used in this review have their characteristics summarised in Table 1.

Table 1. *Characteristics of Studies Included in the Systematic Review*

Ref	Author	Year	Dataset	AI Model	Explainability Method
[5]	Taghipour & Ng	2016	ASAP dataset	Neural Network	Feature importance
[6]	Dong et al.	2017	ASAP dataset	RCNN	Attention visualization
[8]	Rudin	2019	AI decision systems	Interpretable ML	Transparent models
[9]	Khosravi et al.	2022	Educational AI systems	XAI framework	Model interpretability
[11]	Ribeiro et al.	2016	ML models	Model-agnostic systems	LIME
[12]	Lundberg & Lee	2017	ML models	Tree/Deep models	SHAP
[13]	Kumar & Boulanger	2020	Student essays	Deep learning AES	SHAP explanations
[14]	Holmes et al.	2019	Educational AI	AI assessment frameworks	Ethical transparency
[15]	Misgna et al.	2024	AES datasets	Deep learning models	Model explainability
[16]	Alikaniotis et al.	2016	ASAP dataset	LSTM	Saliency analysis
[17]	Cummins et al.	2016	Essay corpus	Regression models	Feature interpretability
[18]	Ke & Ng	2019	ASAP dataset	Neural AES	Linguistic feature analysis
[19]	Mathias & Bhattacharyya	2018	Essay datasets	Neural AES	Feature attention
[20]	Farag et al.	2018	Educational corpus	LSTM	Attention-based explanation
[21]	Jin et al.	2020	ASAP dataset	Transformer	Interpretability analysis
[22]	Mizumoto & Eguchi	2023	English essays	GPT-based scoring	Feature explanation
[23]	Villegas-Ch et al.	2023	Moodle learning data	XAI educational assistant	SHAP

3.2 AI Models Used in Automated Essay Scoring

The review of the studies included in the analysis shows that there is a great variety of machine learning and deep learning models used to solve automated essay scoring tasks. Previous studies were largely based on regression models

and manual generated linguistic features that were good at giving interpretable information about the quality of the essays but lacked the rich ability to represent the complex semantic relationships. In more recent research, deep learning architectures such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs) and convolutional neural networks (CNNs) have gained more widespread use. The purpose of these models is to facilitate automated systems that are able to engage with contextual patterns in writing, enhance the accuracy of scoring as well as allow the assessment of more advanced linguistic attributes like coherence and argument structure. Language models using transformers have also become a topical subject of recent AES studies. Research with models that include contextual embedding models like BERT and other transformer architectures have shown this to be improving by a large margin in terms of predictive performance. Nevertheless, the sophisticated models tend to be black-box structures and not easily analyzed by the educators to understand the logic behind such automated scoring processes. This has led to a growing interest of researchers in the need to incorporate methods of explainability as well as models of predictive performance.

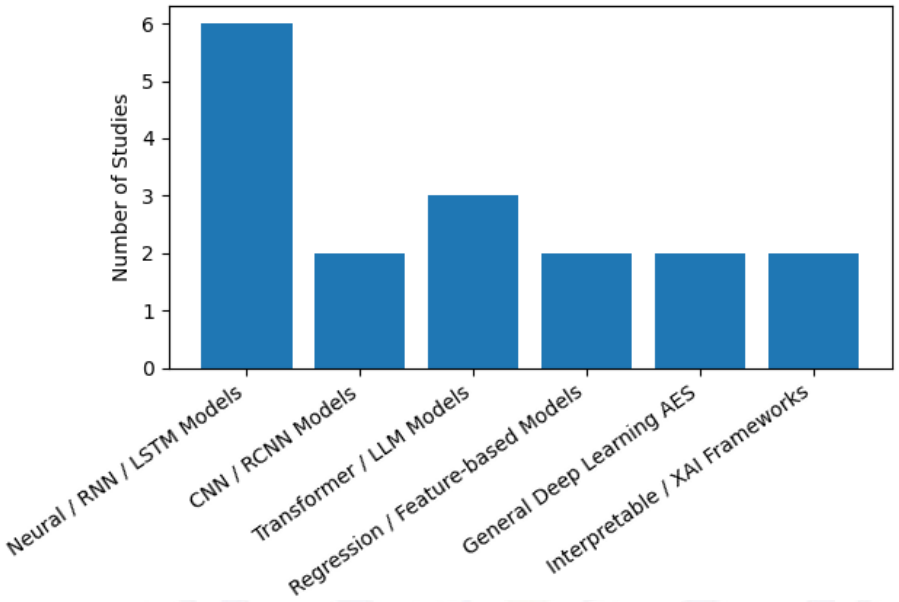

Figure 2. *Distribution of AI Models Used in Automated Essay Scoring Studies*

3.3 Explainable AI Techniques Applied in AES Systems

The techniques of explainable AI were applied to the included studies on a large scale to achieve transparency in automated essay scoring models. One of the most commonly used techniques was model-agnostic interpretability tools like LIME and SHAP, allowing a researcher to determine the importance of specific traits of text on model prediction. Such methods enable teachers to determine what particular features of an essay, including the presence of lexical diversity, sentence structure, or semantic relevance, contribute to the decision-making process of automated scoring. Neural networks that are based on attention were also widely used to bring out important words or sentences in essays. Attention visualization gives an explanation by showing the parts of the text that were most crucial in the process of evaluating the model.

This method is more convenient with deep learning models because it enables researchers to make sense of how the model processes linguistic information over longer texts. Besides the methods of feature attribution, certain studies suggested human-centric explainability models that were aimed at enhancing the educator-computerized scoring

systems interaction. These models propose to provide educational frameworks where model description is easily comprehensible in terms of understanding to educators without technical literacy to enhance trust and human-AI cooperation in teaching assessment contexts.

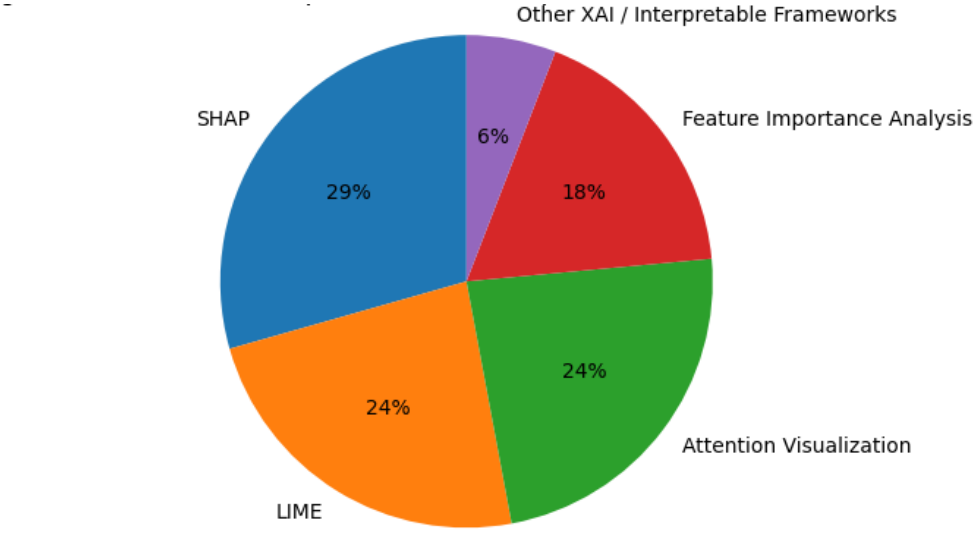

Figure 3. *Distribution of Explainable AI Techniques Used in AES Research*

3.4 Trends in Explainable AES Research

The time distribution of the studies included demonstrates that the research work on the topic of explainable AI in automated essay scoring has significantly risen since 2019. Previous studies in AES were mainly oriented on enhancing the accuracy of the scoring and the performance of the model, but the recent studies are more concerned with the aspects of transparency, fairness, and ethical issues related to the assessment systems based on AI. This change is an indication of the trends in the body of responsible AI and increased consideration of the relevance of interpretability in high-stakes decision-making processes.

Moreover, a number of studies emphasized the role of educator views in the development of explainable AES systems. Researchers have proposed that clear scoring models should be used to assist teachers to authenticate algorithmic responses, detect possible biases, and offer valuable feedback to learners. As a result, explainable AES models are now seen as automated grading tools, but also decision-support systems to help educators to assess student writing.

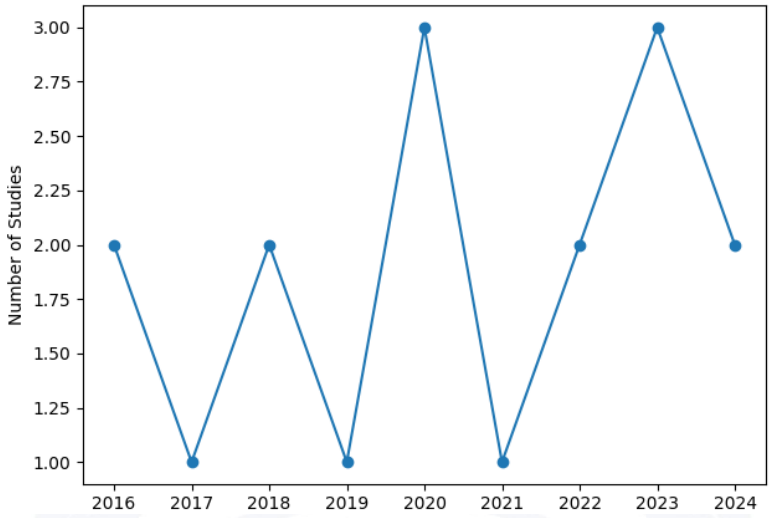

Figure 4. *Publication Trend of Explainable AES Studies (2016–2024)*

Based on the results of this systematic review, explainable artificial intelligence is emerging as a significant part of automated essay scoring studies. In all the 17 studies, the researchers emphasized the need to incorporate explainability methods to improve transparency, interpretability, and trust of AI-based educational assessment systems. Although modern deep learning-based models have enabled the study of scoring accuracy to a high level, the black-box character implies that one has to resort to the explainability tools that will disclose the characteristics behind the automated scoring decisions. Altogether, the literature suggests that integrating high-performing AI models with interpretable explanation mechanisms will provide an encouraging avenue of creating transparent and trustful automated essay scoring systems. Nonetheless, no additional studies address the effectiveness of explainable AES systems in a real educational setting and how educators and students will respond to algorithmic explanations when assessing a student.

DISCUSSION

The current systematic review was designed to summarize the current body of knowledge regarding the concept of integrating explainable artificial intelligence (XAI) methods into automated essay scoring (AES) systems. Through the analysis of 17 publications in the period between 2016 and 2024, one can note that the intersection between machine learning, natural language processing and educational assessment is increasingly growing. The findings show that even though the automated essay scoring systems have already achieved considerable improvements in the predictive performance with the assistance of advanced machine learning systems, the concerns of transparency and interpretability are also the critical factors of its acceptability within the academic context. The harmonization of explainable AI practices has therefore turned out to be a major research agenda in improving the transparency, accountability and pedagogical utility of automated assessment systems. The increased application of deep learning models to automated assessment of essays can be listed among the most significant findings of this review. The core of the early AES strategies was also regression-based models and manually constructed linguistic features that did not have enough explanatory power but offered a certain level of interpretability. More recent studies have adopted the neural network architecture such as recurrent neural networks, long short-term memory networks and transformer-based models to examine the contextual dependence in text and evaluate the high-order writing behavior of the writing such as coherence, quality of argument and topic relevance [16].

These models have been proven to be good predictors and have good concord with human raters hence those of interest when it comes to conducting large scale educational assessment. The more complex models have caused some concern over their interpretability though, as they have become more complex. Deep learning systems are usually black-box

systems where the decision-making process of the systems is not easily decipherable even to the scholars who developed them. This invisibility carries a great implication in the aspect of fairness, accountability and fidelity, which is a big problem in the field of education. The implications of the assessment findings on student academic performance and institutional choices are extremely grave, and that suggests that the stakeholders ought to be in a position to understand how the automated systems can generate the evaluation results. Previous literature has highlighted that transparency is an exceptionally salient requirement in responsible application of AI in stakes high decision making environments [19].

Explainable artificial intelligence methods have thus been proposed as a tool of enhancing the interpretability of automated essay scoring models. The papers, involved in the review, support the statement that some of the explainability methods have been implemented in the AES research, among which are the feature importance analysis, visualization of attention, or model-agnostic techniques such as LIME and SHAP. The methods help to understand which factors affect model predictions by determining which textual features or language patterns lead to scoring decisions.

In case of attention-based neural models, it is possible to highlight words or sentences that the model finds especially relevant during assessment of essay quality. Equally, SHAP-based analysis enables scholars to determine the frequency of the contribution of particular characteristics to the overall score produced by a machine learning model. Explainability procedures in AES systems present a number of possible advantages of educational assessment. First, clear models can increase the faith of educators as they will be able to learn and confirm the decisions of automated scoring. The more teachers can see the logic of the algorithmic predictions, the more they will consider automated systems as aids and not alternatives to human judgment. Second, explainable AES systems can offer quality feedback to learners by showing them what their writing is strong and weak, and this can facilitate the process of formative learning. Instead of just giving a numerical rate, explainable systems can provide particular areas about writing that affected the score e.g. vocabulary use, grammatical correctness, and coherence. Such feedback can help students learn better about the quality of writing and achieve better learning results.

The other significant insight of this review is the increased interest in human-centered AI solutions in the field of educational technology studies. Multiple sources emphasized the necessity of creating explainability mechanisms that can be interpreted and helpful to educators that might lack any technical knowledge in machine learning [21]. This view is indicative of a wider trend in AI research of building systems to aid human-AI cooperation, as opposed to making fully automated choices. Automated essay scoring systems ought to be considered, in the context of education, as decision-support systems that would help educators to assess the work of students and leave the aspects of human control and pedagogical discretion. Even with such encouraging changes, there are a number of research gaps. To start with, most of the reviewed studies were mainly concerned with the development of technical models and experimental analysis by using benchmark datasets, including the Automated Student Assessment Prize (ASAP) dataset.

Although these datasets can be a great resource in the development of algorithms, they might not be the most realistic representation of the real world environment in education. The further development of the explainable AES systems usage in the practically-authentic classroom should thus be studied in the future to assess the interaction between educators and students with the algorithmic explanations in the field. Second, the number of studies that consider the effects of explainable AES systems on instructional practices and learning outcomes is small. Recognizing the impact of explainability on teacher trust, student engagement, and writing improvement is a significant future study direction.

There are also some limitations of this review. The target of the study was mostly peer-reviewed English-language publications included in the reference lists of big academic databases, which might have led to the omission of other valuable studies published in different languages or new outlets. Moreover, because the methodologies of the involved studies varied, the analysis was done based mostly on qualitative synthesis, but not quantitative meta-analysis. The

review, however, gives a detailed picture of the research trends going on in explainable automated essay scoring and identifies the directions that researchers should take in the future.

In general, the results of this systematic review indicate that a fruitful avenue to create transparent, trustworthy, and educationally relevant AI based assessment instruments is the incorporation of explainable artificial intelligence methods in automated essay marking systems. Since educational institutions are progressively embracing AI technologies, transparency and accountability in automated decision-making will be crucial in ensuring that there is fairness and trust in digital learning institutions.

CONCLUSION

The systematic review has explored the present state of explainable artificial intelligence (XAI) in automated essay scoring (AES) systems by summarizing the results of 17 papers published since 2016. The findings reveal that, although machine learning and deep learning developments have enhancing is highly accurate and efficient in automated scoring of essay, the issues of transparency, interpretability, and trust would be relevant challenges in implementing these systems in education. Explainable AI methods like feature attribution, attention visualization, and model-agnostic interpretability tools have shown significant potential in ensuring that these issues are tackled, as they provide information on how automated scoring models process written answers. Through exposing the language and semantic characteristics affecting the algorithmic predictions, explainable AES systems can be used to improve transparency, educator trust, and be used to contribute to more responsible deployment of AI-assessment technologies.

The further research directions should be directed at the further development of the real-world usage of explainable AES systems and the investigation of the way educators and students may respond to algorithmic explanations in classrooms. Specifically, the research is required to assess the effectiveness of explainable scoring systems in enhancing the learning process, promoting formative feedback, and increasing the confidence of a teacher in AI-based assessment instruments. Further studies are also needed to examine multilingual and other cross-disciplinary essay data to enhance the transferability of AES models to a wider range of educational settings. Also, incorporation of human-oriented design and ethical AI models will be necessary in the effort to make sure that automated assessment technologies are transparent, fair, and, at the same time, pedagogical. The future research can further help the creation of credible AI systems that can be used alongside human judgment to improve the overall nature of the educational assessment.

REFERENCES

1. Dikli S. Automated essay scoring. Turk Online J Distance Educ. 2006;7(1):49–62.
2. Page EB. Project Essay Grade: PEG. In: Shermis MD, Burstein J, editors. Automated Essay Scoring: A Cross-disciplinary Perspective. Mahwah (NJ): Lawrence Erlbaum Associates; 2003.
3. Attali Y, Burstein J. Automated essay scoring with e-rater® V.2. J Technol Learn Assess. 2006;4(3).
4. Shermis MD, Hamner B. Contrasting state-of-the-art automated scoring of essays. In: Shermis MD, Burstein J, editors. Handbook of Automated Essay Evaluation. New York: Routledge; 2013. p. 313–346.
5. Taghipour K, Ng HT. A neural approach to automated essay scoring. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP). Austin: Association for Computational Linguistics; 2016. p. 1882–1891.
6. Dong F, Zhang Y, Yang J. Attention-based recurrent convolutional neural network for automated essay scoring. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL). Vancouver: Association for Computational Linguistics; 2017. p. 1537–1546.
7. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). Minneapolis: Association for Computational Linguistics; 2019.

8. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206–215.
9. Khosravi H, Shum SB, Chen G. Explainable artificial intelligence in education. Comput Educ Artif Intell. 2022;3:100074.
10. Gunning D, Aha D. DARPA's explainable artificial intelligence program. AI Mag. 2019;40(2):44–58.
11. Ribeiro MT, Singh S, Guestrin C. 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. San Francisco: ACM; 2016. p. 1135–1144.
12. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS). Long Beach: NeurIPS; 2017. p. 4765–4774.
13. Kumar V, Boulanger D. Explainable automated essay scoring: Deep learning really has pedagogical value. Front Educ. 2020;5:572367.
14. Holmes W, Bialik M, Fadel C. Artificial intelligence in education: Promises and implications for teaching and learning. Boston: Center for Curriculum Redesign; 2019.
15. Misgna H, Singh A, Saha S. Deep learning approaches for automated essay scoring: A systematic review. Artif Intell Rev. 2024;57:45.
16. Alikaniotis D, Yannakoudakis H, Rei M. Automatic text scoring using neural networks. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL). Berlin: Association for Computational Linguistics; 2016. p. 715–725.
17. Cummins R, Zhang M, Briscoe T. Constrained multi-task learning for automated essay scoring. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL). Berlin: Association for Computational Linguistics; 2016. p. 789–799.
18. Ke Z, Ng V. Automated essay scoring: A survey of the state of the art. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI). Macao: IJCAI; 2019. p. 6300–6308.
19. Mathias S, Bhattacharyya P. Automated essay scoring with recurrent neural networks. In: Proceedings of the 27th International Conference on Computational Linguistics (COLING). Santa Fe: COLING; 2018. p. 358–369.
20. Farag Y, Yannakoudakis H, Briscoe T. Neural automated essay scoring and coherence modeling for academic writing. In: Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications. New Orleans: Association for Computational Linguistics; 2018. p. 102–111.
21. Jin C, He B, Hui K, Sun L. TDNN: A two-stage deep neural network for automated essay scoring. IEEE Trans Learn Technol. 2020;13(4):743–755.
22. Mizumoto A, Eguchi M. Exploring the potential of using an AI language model for automated essay scoring. Comput Educ Artif Intell. 2023;4:100131.
23. Villegas-Ch W, Arias-Navarrete A, Palacios-Pacheco X. Proposal of an explainable artificial intelligence educational assistant for learning environments. Int J Educ Technol High Educ. 2023;20:25.