IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 April 2024, accepted 5 May 2024, date of publication 9 May 2024, date of current version 7 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3399163

TOPICAL REVIEW

A Systematic Literature Review: Are Automated Essay Scoring Systems Competent in Real-Life Education Scenarios?

WENBO XU[1,2], ROHANA MAHMUD[1], AND WAI LAM HOO[1], (Member, IEEE)
[1]Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia
[2]Faculty of Digital Arts, Jiangxi Arts and Ceramics Technology Institute, Jingdezhen 333001, China

Corresponding authors: Rohana Mahmud (rohanamahmud@um.edu.my) and Wai Lam Hoo (wlhoo@um.edu.my)

This work was supported in part by Ministry of Higher Education under Fundamental Research Grant Scheme under Grant FP054-2019A, in part by Universiti Malaya (UM) Research Maintainance Fee under Grant RMF1510-2021, in part by the National Natural Science Foundation of China under Grant 32260154, and in part by China-Montenegro Intergovernmental Science and Technology (S&T) Cooperation under Grant 2018-3-3.

ABSTRACT Artificial intelligence technology is becoming increasingly essential to education. The outbreak of COVID-19 in recent years has led many schools to launch online education. Automated online assessments have become a hot topic of interest, and an increasing number of researchers are studying Automated Essay Scoring (AES). This work seeks to summarise the characteristics of current AES systems used in English writing assessment, identify their strengths and weaknesses, and finally, analyse the limits of recent studies and research trends. Search strings were used to retrieve papers on AES systems from 2018 to 2023 from four databases, 104 of which were chosen to be potential to address the posed research aims after study selection and quality evaluation. It is concluded that the existing AES systems, although achieving good results in terms of accuracy in specific contexts, are unable to meet the needs of teachers and students in real teaching scenarios. The improvements of these systems relate to the scalability of the system for assessing different topics or styles of the essays, the accuracy of the model's predicted scores, as well as the reliability of outcomes: improving the robustness of AES models with some adversarial inputs, the richness of AES system functionality, and the development of AES assist tools.

INDEX TERMS Automated essay scoring, deep learning, machine learning, natural language processing, writing assessment.

I. INTRODUCTION

Nowadays, one urgent issue is relieving teachers' burden from the assessment of essays, especially in English standardisation examinations, as they have to spend loads of time and energy evaluating the candidates' writing ability in significant cases. It is unfortunate but true that examiners find it hard to stay consistent in scoring the writing tasks, mainly because, nowadays, composition evaluation methodology primarily relies on humans, even though they have globally unified scoring criteria to follow. Chances are that reviewers' subjective preferences, physical exhaustion, emotion, and other factors might all easily skew the results [1]. Thus, some automated scoring systems have been developed to cope with this challenge.

Regularly assessing writing abilities has made the development of automated text quality analysis technologies crucial for institutions. The amount of time and work required to score the enormous amount of text data produced is too considerable due to the high degree of engagement for human evaluators to do practically [2]. Manual assessment systems involving several evaluators may be susceptible to erroneous judgment due to evaluator conflicts. Creating a system that allows essays to be automatically evaluated with minimal human interaction seems to be the ideal approach to satisfy the expanding needs of the education industry while lowering

The associate editor coordinating the review of this manuscript and approving it for publication was James Harland.

inter-evaluator disagreements. Multiple groups have focused their research on Automated Essay Scoring (AES[1]) systems to combat the difficulties mentioned above [3].

In the past decades, many researchers have worked on the automated scoring of essays. Compared to short answer questions, essay questions are more open-ended in their answers. Students' papers are usually short essays of a few hundred or even thousands of words consisting of a certain number of paragraphs with different topics and writing styles. Students are asked to explain a particular topic, develop an argument and analysis around their ideas, and support their views with their experience. These requirements make the essay assessment difficult, and developing such an evaluation system is even more challenging. It has proven to be quite difficult to evaluate an essay based on all of the factors listed above, including how the essay's content relates to the prompt, the way ideas are developed and articulated, the essay's coherence and cohesiveness, its lexical resources, and its grammatical variety and accuracy [4].

Our research engages with a rich body of existing scholarship on Automated Essay Scoring (AES) systems, recognizing the multitude of literature reviews, dissertations, and publications that have contributed to understanding their application and evolution globally. In this context, our study offers a comprehensive review tailored to address unique research questions relevant to our study's focus. Guided by a systematic review methodology inspired by Keele, our study meticulously follows a predefined protocol to facilitate an exhaustive and impartial analysis [5].

Initially, we outlined the study's objectives, which informed the identification of key research themes central to our inquiry. Subsequently, a carefully structured search strategy was devised, incorporating specific study questions and predefined search keywords across relevant literature sources. Rigorous inclusion and exclusion criteria were applied to select studies closely aligned with our research objectives. To ensure the reliability and rigour of the selected studies, we developed and implemented a robust quality assessment checklist. Finally, a comprehensive analysis of the scrutinized papers was conducted, extracting pertinent insights and findings crucial to advancing scholarly understanding in the field of AES. This methodological rigour sheds light on current advancements within the field and identifies critical research gaps, thus providing valuable insights to guide future exploration.

The paper's structure is organized as follows: Section II provides an overview of related work on AES systems. Section III details the study methodology, elucidating the systematic review approach employed. In Section IV, the results of the systematic literature review (SLR) data analysis are presented, utilizing the suggested technique. Section V discusses considerations arising from the SLR results and outlines the review's limitations. Finally, Section VI concludes the research, summarizing key findings and offering reflections on potential avenues for future exploration.

[1]AES refers as automatic essay scoring in some papers. However, it is more commonly recognised as automated essay scoring. So in this paper, AES is referred as automated essay scoring.

II. RELATED WORK

A small number of systematic literature reviews regarding AES in recent years have been found. Some representative works have been selected and presented in this section. The aspects covered in these reviews are presented in the following table 1.

Hussein et al. conducted a comprehensive review of literature on automated and handmade features in Automated Essay Scoring (AES) [6]. They compared seven types of handmade AES systems, including Project Essay Grader™ (PEG), Intelligent Essay Assessor™ (IEA), E-rater®, Criterion℠, IntelliMetric™, MY Access!, and Bayesian Essay Test Scoring System™ (BETSY), with four types of automatic AES systems based on neural networks. While deep learning-based techniques have shown better performance compared to earlier methods, they may not necessarily yield superior results when modeling complex linguistic and cognitive qualities inherent in essays.

Chassab et al. introduced a novel taxonomy for AES feature analysis, categorizing AES techniques into supervised and unsupervised paradigms [7]. This paper summarizes existing research in AES and highlights three significant limitations of these techniques. Additionally, Ramesh and Sanampudi and colleagues conducted a SLR on AES systems using machine learning (ML) techniques [4]. Their study aimed to identify available datasets for AES research, extracted features for essay assessment, explored the application of ML approaches, and proposed methods to assess the correctness of AES systems. Their analysis of 62 screened articles revealed gaps such as the absence of mechanisms to determine essay completion, lack of AES systems responding to student feedback, and the need for domain-specific essay datasets for training and testing AES systems.

Lim et al. developed three supervised generic architectures for AES: content similarity, ML, and hybrid approaches [8]. They introduced a novel framework for assessing essays based on content and linguistic characteristics, and proposed five evaluation methods to assess the effectiveness of AES models, advocating for the use of quadratic weighted kappa (QWK) as a common technique. Furthermore, Morade and Netak summarized three essential data engineering tasks in preprocessing data for AES: feature engineering, word embedding, and measuring text similarity. They presented and compared several common methods for Automated Grading of Essays (AEG) in their systematic literature review (SLR) [9].

Zhang and Liu reviewed AES models, which, from the standpoint of developing natural language processing (NLP) approaches, solely rely on deep neural networks (DNNs). They illustrated the principal concept and specific design of each existing representative DNNs-AES model [10]. Similarly, Uto offered a detailed introduction to each model

TABLE 1. Summary of related literature reviews in the field of AES.

Authors [Ref.]	Year	Technique(s)	Feature(s)	Dataset(s)	Metric(s)	Limitation(s)
Hussein et al. [6]	2019	Yes	Yes	No	Yes	Dataset utilized for model training has not been comprehensively incorporated
Chassab et al. [7]	2021	Yes	Yes	Yes	Yes	Latest models not covered; insufficient articles selected
Ramesh and Suresh [4]	2021	Yes	Yes	Yes	Yes	Insufficient articles selected; no prediction on future trends
Chun Lim et al. [8]	2021	Yes	Yes	No	Yes	Dataset utilized for model training has not been covered
Borade and Netak [9]	2021	Yes	Yes	No	Yes	Dataset for model training not covered; methods employed by models not sufficiently summarized
Zhang, J.and Liu, J. [10]	2022	Yes	Yes	No	Yes	Only neural network-related models were summarized; Dataset utilized for model training has not been covered
Masaki Uto [11]	2021	Yes	Yes	No	Yes	Only neural network-related models were summarized; Dataset utilized for model training has not been covered

and a thorough review of DNN-AES models by classifying them into four categories: prompt-specific holistic scoring, prompt-specific trait scoring, and cross-prompt holistic scoring, and cross-prompt trait scoring [11].

It is worth noting that since its emergence in 2023, ChatGPT has been influential in AI and continually transformed automated scoring techniques. This review covers specific contents of articles published in 2023, which are closely related to the latest modeling techniques, to provide a more comprehensive overview of the techniques. Additionally, further refinement and summarization of more articles are required to enhance the review of other aspects of the AES field: datasets, features, and evaluation methods.

III. METHODOLOGY

The procedures necessary to conduct a thorough systematic review are outlined in a protocol inspired by Keele et al., which prevents researcher biases in a predetermine manner. In the initial stage, the purpose of this study served as the basis for several research topics [5]. The search strategy was implemented afterward, using the specified study questions, and it involved the selection of search keywords and literature resources. After that, they were selected based on inclusion and exclusion criteria to ensure the studies were relevant. A quality assessment checklist was designed to assess these relevant studies to guarantee their quality. Finally, all the scrutinised papers were analysed.

A. RESEARCH QUESTIONS

This SLR aims to find why AES has failed to gain widespread popularity in real-life educational scenarios and to summarise the state-of-the-art in the field to provide ideas and directions for further research. Four research questions (RQs) were developed in order to achieve this aim, and they are mentioned below:

RQ1:What are the currently existing essay assessment systems and AI technologies apply to English writing assessment?

This question can be answered by making a table to visually list the currently developed essay assessment systems, including their years of development, areas of applicability or exam categories, and the technologies they applied.

RQ2:What features have been extracted from the essay for English writing assessment?

The response to the question can provide information about the different features that have already been extracted and the libraries that were utilised.

RQ3:Which evaluation metrics and datasets have been applied to evaluate the accuracy and dependability of these technologies and systems?

To determine if new approaches require creation or improvement, a list of the most widely used datasets and assessment metrics may be compiled.

RQ4: What are the challenges or limitations of the currently developed systems and technologies?

The answer to this question is to allow subsequent researchers to build on the system's performance and to point out future research directions for subsequent research.

B. SEARCH STRATEGY

The following is a thorough summary of the search tactics used in this study, including the search terms, literature sources, and search procedure.

1) SEARCH TERMS

There are mainly 3 ways to find suitable search terms for our study:

- Significant terms come from the questions posed by the research;
- For key phrases, alternate spellings and synonyms are provided;
- Major keywords are found in relevant papers or books.

Then these major terms are linked together using the Boolean AND and OR operators. Ifenthaler suggests the search strings this paper used as follows: Automatic AND (''essay''/OR ''writing'') AND (''scoring''/ OR ''grading''/OR ''evaluation'') [12]. Whatever search strings are used, they are uniformly referred as AES in this paper.

2) LITERATURE SOURCES

To find suitable data for our SLR, we automatically searched popular computer science databases like Scopus, Google Scholar, IEEE Explore, and Web of Science.There are many prospective journals, conference proceedings, workshops, symposiums, book chapters, and IEEE bulletins in these well-known digital databases mentioned above. If search terms can be effectively applied while searching for article titles, abstracts, and keywords, then a comprehensive range of relevant articles can be efficiently filtered for our SLR.

3) SEARCH PROCEDURE

An SLR involves thoroughly searching all relevant materials related to a discussion topic.The following stages of the search procedures applied to this research, as seen in Figure 1.

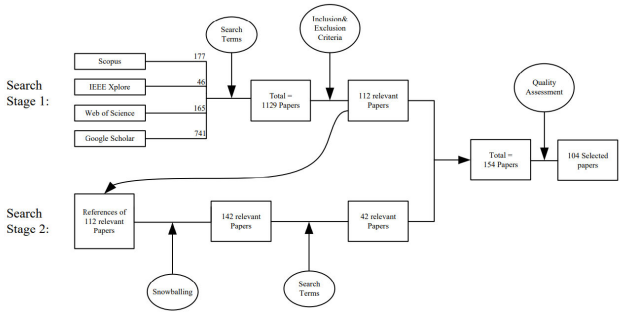

FIGURE 1. Two stages of the search procedure.

C. PAPER SELECTION

In the first search stage, 1129 papers were retrieved from four databases by our search terms, with 177 papers in Scopus, 46 papers in IEEE Xplore, 165 papers in Web of Science and 741 papers in Google Scholar, respectively. Also, 112 pertinent papers were chosen after the titles and abstracts of the 1129 publications were filtered using the inclusion and exclusion criteria. In the second search stage, the references of the selected 112 papers were examined through a process known as "snowballing[2]" to find any significant studies that could have been overlooked during the original search. As a result, 142 more relevant papers were found, selected by the same filter as in the previous stage. This endeavour resulted in the discovery of 42 more relevant papers, bringing the total number of chosen studies to 154. Finally, these 154 research were subjected to quality assessment procedures. After the exercise, 104 studies were selected as having the potential to address the posed research questions.

1) INCLUSION AND EXCLUSION CRITERIA

Figure 1 illustrates the 1129 prospective studies discovered during the first search. Therefore, it took great thought to limit this research to pertinent ones. Each study's titles were considered before a cursory review of its contents. As a result, any publications that failed to address the topic under discussion or provide a response to any of the suggested study questions were eliminated from the list of pertinent studies. Furthermore, only research written and published in English was considered for the list of relevant studies. However, when many copies of the same document appear, the most complete, up-to-date, and improved copy is included in the search process, and the others are removed. We conducted a systematic literature assessment on publications published between January 1, 2018, and December 31, 2023, looking for needs-prioritising methods. If the year scope of research is too broad, the number of relevant articles in the search is enormous. Secondly, AI technology changes rapidly, and previous methods can easily be obsolete by the new technology coming out later. Finally, we could also find previously important technical backgrounds in relevant research from recent years. We only targeted relevant articles within the last five years for our SLR study. Table 2 provides an overview of the used examination criteria.

2) QUALITY ASSESSMENT

To assure the quality of the 154 relevant papers, we evaluated each manuscript using quality assessment questions besides the inclusion and exclusion standards. The chosen studies' quality assessment was accomplished using a weighting or scoring approach to identify relevant studies that could respond to each research question. The publications that clearly explained the technique, validation, and outcome analysis were included. The guidelines provided by [5] served as the basis for the format of the quality checklist questions. We developed a set of quality evaluation questions to analyse the chosen studies' veracity, comprehensiveness, and applicability. Table 3 displays these questions. There are just three possible responses to each question: "Yes," "partially," or "No." These three responses are given a score of 1, 0.5, and 0, respectively: "Yes," "Partly," and "No." The study's final score ranges from 0 to 4.

These metrics assessed the reliability of the inferences made and provided guidance for interpreting the results of the chosen research. Additionally, it assisted in determining the validity and coherence of the findings.

D. DATA SYNTHESIS

Data synthesis seeks to address or answer the research questions by consolidating the information from the selected studies. In order to synthesise the data, these 104 chosen papers were further examined to evaluate each study's particular contents about the standards listed in Table 4. In order to increase clarity, a few research will be synchronised. It will also make locating precise answers to the research queries simpler. The research's extracted data contained quantitative and qualitative information (e.g., strengths and shortcomings of current AES approaches), such as the accuracy or result

[2]The "snowballing" method represents an efficacious approach for identifying additional pertinent literature by scrutinizing citations within established literature. It is particularly apt for uncovering and accessing scholarly literature pertaining to a designated subject matter.

TABLE 2. Inclusion and exclusion criteria.

Inclusion Criteria	Exclusion Criteria
a. All papers published in English language	a. Papers that are not published in English language
b. Relevant papers that are published from 2018 to 2023	b. Papers that have little to do with the research question
c. All published papers that have the potential of answering at least one research question	c. Duplicate papers (only the most complete, recent and improved one is included, the rest are excluded.)
d. Papers focuses on Automated Essay Scoring systems	d. Gray papers; i.e. papers without bibliographic information such as publication date/type, volume and issue numbers, were excluded

figures for the AES system. The following is a collection of the thorough arguments for the data synthesis procedures used:

RQ1-related data were arranged logically. The various AES approaches were distributed using visualisation tools like pie charts and bar charts. The taxonomies of characteristics used in RQ2 were identified from a few chosen studies and taken from the assessment essay. The result was recognised and shown via a tabular descriptive diagram. The metrics and datasets utilised in the different AES models were tabulated and presented in RQ3. Simultaneously, some research was used to identify the difficulties or constraints of current methods, and the results were presented in tabular form in RQ4.

TABLE 3. Quality assessment questions.

No.	Questions
1	Are the aims of the research clearly articulated?
2	Is the proposed technique clearly described?
3	Is the experimental design appropriate?
4	Is the experiment applied to adequate project data sets?
5	Does the research add value to the academia or industrial community?

E. THREATS TO VALIDITY

The three main factors that are working against the review protocol (Figure 1) are publication bias, missing important studies, and incorrect data extraction. In this section, the reasons why these three factors could threaten the review protocol are explained, and the methods to address these problems are presented.

Publication bias is the first threat believed to exert an adversarial impact on our SLR. The issue of favourable outcomes being published more frequently than negative results is called publication bias. The idea of positive or bad outcomes might occasionally rely on the researcher's point of view [5]. Researchers are often likely to publish more positive than negative outcomes regarding AES technology or assert that their technique is superior to other technologies, especially when influential groups in the software industry sponsor their methods or techniques. This may lead to the fact that the performance of current AES technology is overestimated. To curb this threat, grey literature and conference proceedings have been scanned; experts and researchers working in the area have been consulted to see whether they are aware of any pending findings. Also, publications comparing the currently available AES technologies were searched for and incorporated into the chosen research to obtain the findings of an objective evaluation of the various AES systems, considering that these comparison studies often give objective findings.

Another potential risk to the validity of our SLR is the omission of significant research. Although we have designed reasonable inclusion criteria and a quality assessment checklist and searched many different digital databases for studies that are highly relevant to the specified research questions in this paper, not all papers can be retrieved using keywords, titles, or abstracts that include terminology relevant to the study topics. To find the studies that were overlooked during the first search, we conducted a thorough manual review of all the retrieved studies' references. The selection criteria were rigorously defined and aligned with the research goals to prevent the unintentional exclusion of desirable studies. The studies were chosen after meticulously applying the quality evaluation criteria, and contradictions were quickly fixed when they appeared. In this manner, a variety of other research was found.

Last but not least, to reduce the risk brought on by incorrect data extraction, all the chosen papers were re-evaluated to find the genuine positives, which are instances when a study's title may suggest its relevance but its contents fail to respond to any of the research questions. The authors independently evaluated the chosen studies using the evaluation questions in Table 3 to limit the inaccuracy of the retrieved data. An inter-rater agreement was processed to settle disagreements and achieve consistency in the author-generated rating order.

IV. RESULTS

A. OVERVIEW OF SELECTED STUDIES

The authors carefully evaluated the quality of the chosen studies. The writers addressed every disagreement over the findings of the quality assessment to come to a consensus. The dependability of the review's conclusions was achieved by only considering relevant studies with acceptable quality rates or with quality scores more prominent than 2 (half of the percentage score). Consequently, 104 final relevant research were chosen after 154 articles from the first gathered studies were eliminated (see Figure 1). The quality ratings of these selected papers are listed in Table A1 as the supplementary material.

A total of one hundred and four research projects were selected for this inquiry. Forty-seven were released in journal articles, forty-six in conference proceedings, nine from

TABLE 4. Contents assessment criteria.

Selected Studies	Description
Identification of study bibliographic references	Unique identification number for the study, publication year, title and source
Type of study	Journal and conference papers, IEEE bulletins and book chapters
Study focus	Domain topic, problems, scope, motivation and objectives
AI technology	Natural Language Processing, Machine Learning, Deep Learning
Data analysis	Quantitative/qualitative analysis
Constraints and future directions	Identification of the study's shortcomings and areas for future research
Application domain	Description of the context and application domain of the study. For example, academic or industrial settings

workshops, and two from symposiums. Figure 2 displays the proportions and corresponding numbers of the selected publications, whereas Figure 3 demonstrates the number of publications by year. However, Table 4 has extensive summaries of the chosen research.

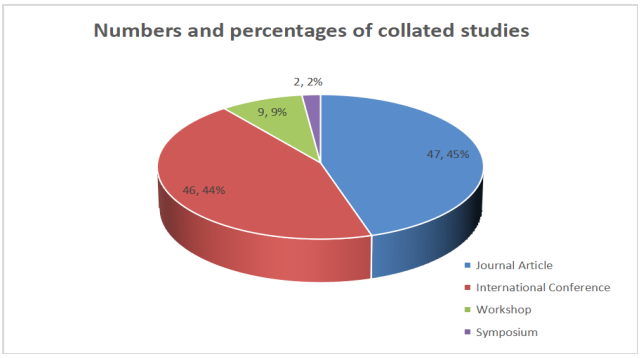

FIGURE 2. Numbers and percentages of collated studies.

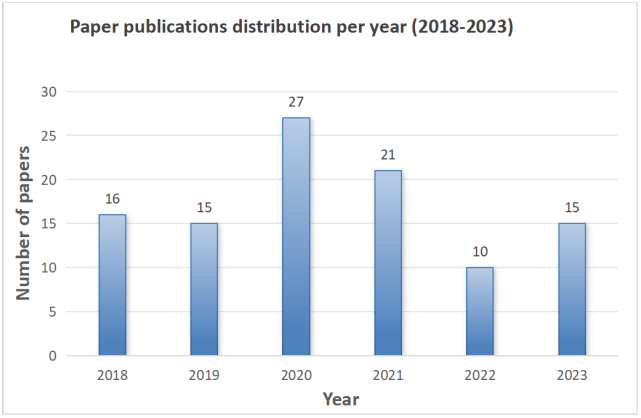

FIGURE 3. Number of publications by year of publication.

B. EXISTING AES MODELS AND AI TECHNOLOGIES (RQ1)

NLP, a subfield of artificial intelligence (AI), aims to improve computers' comprehension and interpretation of natural language spoken by humans through their interactions with devices. Many applications, including sentiment analysis, speech recognition, machine translation, and automatic essay scoring, employ NLP. The mapping of different AI technologies used in existing AES models is displayed in Figure 4. Specific AI technologies and algorithms used in the screened papers is listed in Table 5. In the field of AES, ML, Deep Learning (DL), and NLP intersect to offer comprehensive solutions for evaluating and scoring essays.

AES and ML algorithms are becoming increasingly popular in education. These models assess student essays and provide feedback on their writing. ML models use algorithms to analyse the text and extract features that may be applied to spot trends and patterns in student composition. The algorithms can then assign scores to the essays and provide feedback on areas that need improvement. In addition, these models can provide personalised feedback and advice to assist learners in developing their writing abilities. Automatic essay scoring and ML models have the potential to revolutionise the way that students are assessed and provide them with the feedback they need to become better writers.

ML models present several strengths in the context of automated essay scoring. Firstly, they demonstrate scalability, efficiently managing large datasets and facilitating the rapid scoring of numerous essays. Additionally, these models offer interpretability, as exemplified by traditional techniques like Decision Trees or Support Vector Machines (SVM), enabling stakeholders to discern the factors influencing scores. Moreover, ML allows for the integration of handcrafted features rooted in linguistic and structural characteristics of essays, enhancing the model's performance.

However, these approaches suffer from certain weaknesses. Notably, they may struggle with limited contextual understanding, particularly in grasping intricate relationships within essays that necessitate comprehension of semantics and higher-order language structures. Furthermore, their ability to generalize to unseen essay topics or domains without extensive retraining or adaptation may be limited. Finally, the success of ML models often hinges on the quality of handcrafted features, which can be labor-intensive to engineer and might fail to capture all pertinent information.

Within the field of artificial intelligence (AI), deep learning (DL) refers to a subset of ML techniques based on artificial neural networks with representation learning. DL models are capable of learning from large volumes of data, whether labeled or unlabeled, structured or unstructured, and are applicable across various learning paradigms, including supervised learning, unsupervised learning, and reinforcement learning. DL algorithms are used in various applications, including automatic essay scoring. DL is used

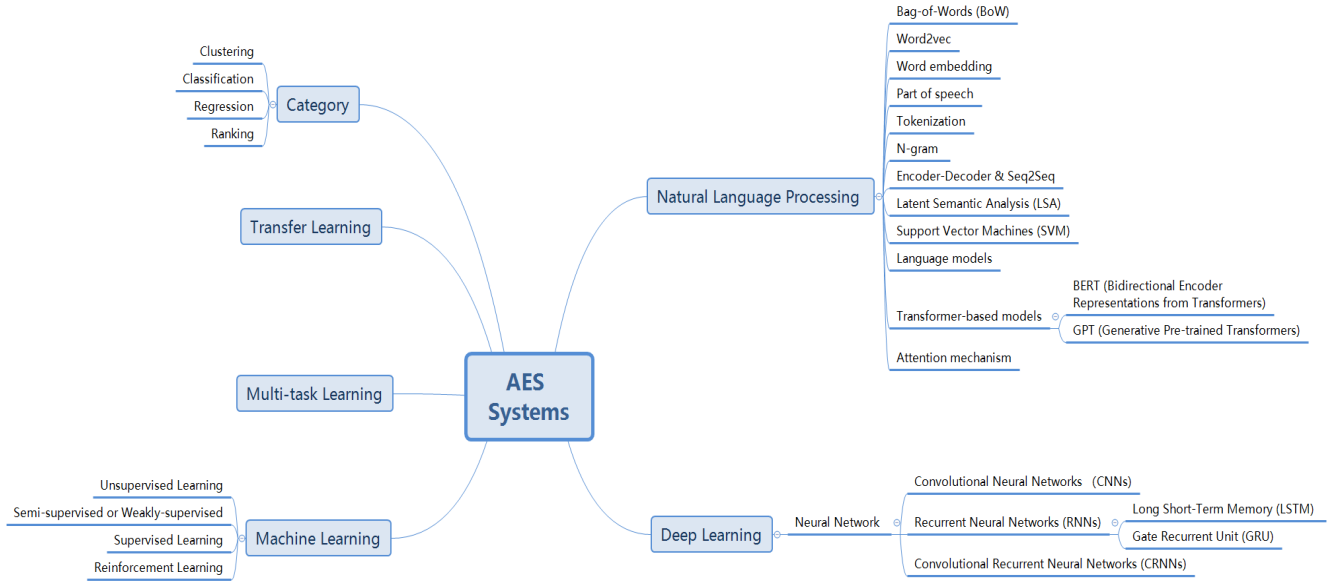

FIGURE 4. The mapping of different AI technologies in AES models.

in automated essay scoring to provide a more comprehensive and accurate analysis of student writing. DL algorithms can understand complex sentence structures and identify relevant topics, which can provide more accurate and meaningful insights about the essay. DL algorithms can also identify patterns in the essay, such as standard errors, which may be utilized to provide the learner feedback. In addition to providing accurate and meaningful insights about student writing, DL can also be applied to enhance the scoring system's precision and uniformity. DL algorithms can be trained to recognise patterns in the essay and adjust the scoring accordingly, ensuring that essays are scored fairly and accurately.

Since the release of Chatgpt came out at the end of 2022, some researchers have discovered the potential of large language models in the AES domain. The efficacy of ChatGPT, a generative artificial intelligence (GAI) model, in essay auto grading was examined by Altamimi [13]. They improve the ChatGPT model for essay grading using the dataset they gathered, fine-tune it, and compare the model's results with essays that have been manually rated. Also, the study conducted by Mizumoto and Eguchi investigated the viability of utilizing GPT for AES [14]. The research problem was tackled by analyzing the correlation between linguistic measures and essay quality, contrasting the method with alternative models, contemplating prompt engineering, deliberating on the constraints and possibilities of GPT-4, scrutinizing ethical and educational implications, and tackling the explainability issue in DL models.

DL models offer notable strengths in automated essay scoring. Firstly, they excel in representation learning, particularly neural networks, which autonomously acquire intricate representations of textual data, thereby capturing complex patterns and dependencies inherent in essays. Moreover, architectures like Recurrent Neural Networks (RNNs) or Transformer-based models facilitate hierarchical feature extraction, enabling a deeper understanding of context within essays. Additionally, the application of transfer learning, leveraging pre-trained language models such as BERT or GPT, allows for effective fine-tuning on essay scoring tasks with minimal data, capitalizing on knowledge gleaned from extensive textual corpora.

However, these approaches are not without weaknesses. DL models often demand substantial annotated data for training, posing challenges in scenarios where such data may be scarce, especially in specialized essay scoring tasks. Furthermore, training these models requires significant computational resources and infrastructure, particularly for large-scale models like GPT. Lastly, the black-box nature of DL models raises concerns about interpretability, hindering stakeholders' ability to understand the rationale behind assigned scores for essays.

In conclusion, the rationale for choosing these models or techniques varies depending on the specific goals of the AES study. Factors such as the complexity of the content of the essay, the availability of annotated training data, computational resources, and the level of interpretability required may influence the selection of the models described above. Researchers typically choose models that meet the requirements of the task and are effective in capturing the nuances of essay content, resulting in more accurate and reliable scoring results.

C. EXTRACTED FEATURES OR VECTORS (RQ2)

The endeavor of feature creation constitutes a significant aspect of the labor invested in AES systems. Researchers have diligently extracted specific features or vectors to enhance the accuracy and performance of their models. A comparative analysis with prior studies reveals notable shifts in the features extracted by researchers, progressing from initial, rudimentary attributes such as word count and essay length to more intricate dimensions, including semantic coherence and argument adequacy.

Integrating statistical and linguistic features is paramount, serving as the cornerstone for the comprehensive analysis and evaluation of essays. Statistical features, characterized by their quantitative and statistical attributes commonly associated with textual content, play a vital role in facilitating the model's understanding and assessment of various dimensions within the composition. These dimensions encompass linguistic complexity, organizational structure, and grammatical accuracy, with standard statistical features including vocabulary statistics, paragraph features, and logical relations.

Conversely, linguistic features stand apart from semantic attributes explored in prior reviews, constituting categories such as lexical features, discourse structure features, and syntactic features. These linguistic attributes contribute significantly to the holistic assessment of essays, providing valuable insights into language usage, coherence, and syntactic intricacies.

Recognition of the significance of these features is foundational to fully harnessing the capabilities of AES systems in accurately evaluating and providing feedback on essays. This understanding underscores the imperative of incorporating statistical and linguistic features into AES methodologies to enhance the efficacy and reliability of automated essay-scoring processes.

In addition to statistical and linguistic features, alternative categories of features are commonly denoted as content features or contextual features. These encompass a broad spectrum of textual attributes, including but not limited to topic relevance, linguistic style, and argumentative support. Each specific feature or vector extracted from the scrutinized literature is meticulously cataloged and presented in Table 5. The taxonomy or categorization of these features referred to during the compilation of the relevant tables, is elucidated in the works of [11] and [15].

Feature extraction techniques are essential for designing AES with feedback, as this strategy has the benefits of being interpretable and comprehensible. However, this strategy often requires considerable technical work to attain high-scoring precision for different articles. Deep neural networks (DNNs) are used in a neural network technique that automatically extracts features to eliminate the requirement for feature engineering. For neural models to perform successfully, they need to be trained on large amounts of labeled data. Although there is enough data to train accurate AES models for English essays, this is different for most natural languages. Even though several neural holistic scoring models have attained cutting-edge outcomes, these models may likely be made even better by adding custom features created via feature engineering. Therefore, it is preferable to consider feature-based and neural techniques complementary rather than rival theories.

D. DATASETS/CORPORA AND EVALUATION METRICS (RQ3)

1) DATASETS/CORPORA

Datasets are needed for training and evaluating AES systems because they provide the necessary data for the system to learn from and accurately assess the written essays. In some studies, researchers refer to the datasets they use as corpora. For convenience, we will refer to this concept uniformly as datasets in this paper. For an AES system to accurately evaluate an essay, it must be trained to recognise patterns in the data and to understand the characteristics of good and bad essays. The datasets used to train and evaluate an AES system must include various essays, including essays of different lengths, different topics, and written by different authors. This will ensure that the system can properly identify and evaluate the various types of essays that it will be asked to assess. Additionally, the datasets must be labelled with the correct assessment scores to enable the system to evaluate the essays it is presented with accurately. The datasets must also be diverse and representative of various topics, authors, and writing styles to ensure that the system can accurately assess a wide range of essays.

Each of the datasets used in the screened papers is displayed in Table 4. Table 5 indicates that Automated Student Assessment Prize (ASAP[4]) was utilized in over half of the investigations as their dataset for training and evaluating AES models. The reason why this database is so widely used is probably due to the Kaggle competition held in 2012, where the organisers made the ASAP database publicly available to facilitate participants to compare the performance of their models. Since then, it has gained popularity as a dataset for holistic grading. Yannakoudakis and Cummins gave a detailed description of this database [15], along with a list of the other four most commonly used databases, namely CLC-FCE,[5] TOEFL11,[6] ICLE,[7] and Argument Annotated Essays (AAE). To avoid repetition, the description of all these five datasets will not be repeated in this paper.

In addition to these common publicly available datasets mentioned above, some researchers have created their own databases to meet specific research purposes in some studies. For example, Ye and Manoharan developed a third dataset that consisted of descriptors linked to computer programming expertise to assess the usefulness of the suggested method for grading computer science courses [16]. Two replies plus a label indicating whether the two responses are synonymous

[4]https://www.kaggle.com/c/asap-aes
[5]Cambridge Learner Corpus First Certificate in English.
[6]Test of English as a Foreign Language.
[7]International Corpus of Learner English.

TABLE 5. Techniques used in existing AES models and their characteristics.

Authors [Ref.]	Technique(s)	Model(s)	Feature(s)	Dataset(s)	Metrics
Gao et al. [32]	DL	Convolutional Neural Networks (CNNs)	Statistical features	Non-/Award-winning images (4800/19,200)	Recall+ Precision+ Accuracy
Chen [33]	NLP+DL	Part-of-Speech Tagging +C/RNN	Word features, Content text features, Nontext features	8 ASAP datasets	Precision+ Accuracy+ Recall+ QWK
Zhao and Chen [34]	DL	Seq2Seq + Bi-LSTM	Semantic features	1200 pieces of CET-4 compositions	F1+ Recall+ Precision
Hsiao and Hung [35]	NLP+DL	Fusion Neural Network	Lexical features,Semantic features	N/A[3]	Accuracy+ Recall+ F1-scores
He [36]	DL	LSTM + Transformer Attention	Semantic features	Chinese Learner English Corpus (CLEC)	F1-measure+ Recall + Precision+ F-score
Xiao et al. [37]	ML	TF-IDF XGBoost	Text features	More than 2000 samples provided by Chinese testing International Co., Ltd	MAE, Root Mean Squared Error (RMSE) + Median
Ye and Manoharan [16]	NLP+ML	BERT, RoBERTa + DeBERTa	Linguistic features	SICK dataset + Microsoft Paraphrase Corpus + 1 self-created corpus	Pearson's coefficients and Mean Square Error (MSEs)+ Accuracy
Elks [38]	DL	Transfer Learning	Manually extracted features	COLA corpus (10,000 sentences)	Mean Absolute Error (MAE)+ Root Mean Square Error (RMSE)
Beseiso et al. [39]	NLP+DL	Transformer-based neural network	Language features	8 ASAP datasets	QWK
Chimingyang [40]	ML	LSTM + Word Embedding	Manually extracted features + Complex Word embedding features	8 ASAP datasets	QWK
Oktaviani et al. [44]	NLP+DL	CNN-LSTM	Text features	SIMPLE-O (2,772 augmented data+testing 43 data)	Accuracy
Balaha and Saafan [42]	NLP	HMB-MMS-EMA	Text features	Quora Questions Pairs dataset +HMB-EMD-v1	Cosine similarity + Pearson's Correlation + RMSE + Accuracy
Das et al. [43]	NLP+DL	FacToGrade (RNNs+LSTM networks)	Handcrafted features + Semantic features	8 ASAP datasets	10-Fold K Cross Validation+ QWK
Wang et al. [44]	NLP +DL	bi-LSTM networks + attention	Semantic features	8 ASAP datasets	QWK
Cong [45]	ML	Data fusion	Text features + Syntactic features	Six samples of the ASAP data set	Accuracy
Ratna et al. [46]	ML	Latent Semantic Analysis (LSA) + Winnowing	Word features	43 students' examination answers (5 questions)	Accuracy
Chen and Zhou [47]	DL	CNN + Ordinal Regression (OR)	Word features	ASAP datasets	Mean Square Error (MSE) +QWK
Wei et al. [48]	NLP+DL	Recurrent Neural Networks (RNNs)	N/A	Articles from the New York Times of 2020	Precision of error detection
Vajjala [49]	NLP	N/A	Linguistic features	TOEFL11 Corpus(12,000 essays)+First Certificate of English (FCE) corpus	Pearson Correlation +Mean Absolute Error (MAE)
Li et al. [50]	NLP	HNN-AES	Linguistic features+ Semantic features+ Structure features	8 ASAP datasets	QWK+ Accuracy + Mean Square Error (MSE)
Beseiso and Alzahrani [51]	DL	BERT Embedding	Manually extracted features + Word embedding features	8 ASAP datasets	Kappa statistics + Accuracy+Mean Squared Error (MSE)

TABLE 5. *(Continued.)* **Techniques used in existing AES models and their characteristics.**

Dasgupta et al. [2]	NLP+DL	Deep Convolution Recurrent Neural Network	Linguistic features+ Psychological features +Lexical features	8 ASAP datasets	QWK
Nicula et al. [52]	NLP+DL	Language Models +Transfer Learning	Lexical, Semantic, Syntactic features+ paragraph quality	MSRP+ ULPC + A small dataset containing paraphrases	F1-measure
Hoblos [53]	ML	Latent Semantic Analysis (LSA) +Latent Dirichlet Allocation (LDA)	Semantic features	Over two semesters with 57 students, 118 essays	N/A
Sharma and Jayagopi [54]	NLP	MDLSTM Model +Word Embeddings	Word features	8 ASAP datasets	QWK + F1-score + Accuracy
Kumar and Boulanger [55]	DL	SHapley Additive Explanations (SHAP)	Lexical+ Semantic+Syntactic features	8 ASAP datasets	QWK + Precision + Recall + F1-scores
Wang et al. [20]	NLP + DL	Bidirectional Encoder Representations Transformers (BERT)	Multi-scale (token-scale, segment-scale and document-scale)features	8 ASAP datasets + CommonLit Readability Prize (CRP2) dataset	QWK + Average + RMSE
Jeon and Strube [56]	NLP+ML	Considering- Content-XLNet	Essay length + Content features	8 ASAP datasets +TOEFL dataset	QWK Average + Accuracy
Cozma et al. [57]	ML	HISK+BOSWE + v-SVR	High-level semantic features	8 ASAP datasets	QWK
Kumar et al. [58]	ML+DL	STL-(Bi)LSTM and MTLLSTM and MTL-BiLSTM	Essay trait features	ASAP ++datasets	QWK+ Five-fold cross validation
Liu et al. [59]	NLP+DL	Feature-engineered + end-to-end model	Handcrafted Features + Semantic features	8 ASAP datasets	QWK
Farag et al. [60]	NLP+DL	Local Coherence (LC) +LSTM AES Model	Connectedness features	8 ASAP datasets	QWK + Total Pairwise Ranking Accuracy (TPRA)
Zhang and Litman [61]	NLP+DL	Co-Attention Based Neural Network model	Word embedding vector+Text features	Four ASAP datasets(Prompts 3,4,5,6)	QWK
Kabra et al. [24]	NLP	General Framework for Test Evaluation Model	N/A	8 ASAP datasets	QWK
Manabe and Hagiwara [30]	NLP	EXPATS-LIT+ASAP-AES Scorer	N/A	ASAP-AES dataset	Precision, recall, and F1 measure+ Accuracy + PCC +QWK
Ludwig et al. [62]	NLP+ML	Transformer Models	N/A	DomPL-IK(780 samples)	QWK+ Accuracy+ ROC AUC+ F1-measure
Jong et al. [63]	NLP+ML	Manipulating-Length-GRU + Considering-Content-LSTM	N/A	8 ASAP datasets	QWK
Ludwig et al. [68]	NLP+DL	Transformer Models with data augmentation	Semantic features	8 ASAP datasets	QWK + Accuracy
Hardy [26]	NLP+DL	PD-Sentence-BERT with multi-head attention	Atomic vectors+Semantic features	8 ASAP datasets	QWK
Ormerod et al. [64]	NLP	N/A	Linguistic features	A corpus of 15,000 single-scored responses	QWK+ Accuracy+ Standardized Mean Difference(SMD)
Ormerod et al. [65]	NLP+DL	Efficient transformer-based language models	N/A	8 ASAP datasets	QWK
Muang kammuen and Fukumoto [66]	NLP+DL	Hierarchical neural network model+Multi-Task Learning (MTL)	Lexicon and text semantic	8 ASAP datasets	QWK + MSE
Uto et al. [67]	NLP+DL	DNN-AES model	Handcrafted essay-level features	8 ASAP datasets	QWK
Yang et al. [68]	NLP+DL	Multi-loss to fine-tune BERT models	Text features +Word features	8 ASAP datasets	QWK
Mim et al. [25]	NLP	An unsupervised pre-training model +Masked Language Modeling (MLM)	Sentence and Discourse Indicator Corruption	8 ASAP datasets	Mean Squared Error (MSE)
Ridley et al. [69]	NLP	Prompt Agnostic Essay Scorer (PAES) model	Non-prompt-specific features + General features	8 ASAP datasets	QWK

TABLE 5. *(Continued.)* **Techniques used in existing AES models and their characteristics.**

Zhang and Litman [70]	NLP+DL	Neural Network	Topical Components (rubric-based features)	2 RTA source texts	QWK + Pearson
Tsai et al. [28]	NLP+DL	RMSProp+CNN +GED	N/A	EF-Cambridge Open Language Database (EFCAMDAT)	QWK
Hellman et al. [27]	NLP+ML	KNN+Multiple Instance Learning (MIL)	N/A	Pearson proprietary corpus	N/A
Mayfield and Black [71]	NLP+DL	Fine-Tune BERT	Semantic features	5 ASAP datasets	QWK
Hirao et al. [17]	NLP+ML+DL	LSTM+BERT model	Holistic+ content + organization + language Features	GoodWriting dataset	Root Mean Squared Error(RMSE) + QWK
Jeronimo et al. [72]	NLP+ML	Computing with Subjectivity Lexicons	Predefined features	1,840 essays were written by high-school students as part of a standardised Brazilian national exam	Area Under Precision-Recall Curve (PR-AUC)+Average ROC-AUC
Rodriguez et al. [19]	NLP+DL	BERT+ XLNet Pytorch-Transformers	Text features	8 ASAP datasets	Accuracy + QWK / Cohen's Kappa Score
Berggren et al. [73]	NLP+DL	GRU-based attention model	N/A	ASK corpus	Macro and micro F1-scores
Nadeem et al. [74]	NLP+ML+DL	Discourse-Aware Neural Models(HAN+BCA)	Contextualised embeddings	ETS Corpus+ASAP datasets 1/2	Ten/Five-fold cross-validation + Accuracy
Carlile et al. [75]	NLP+ML	Novel computational models	N/A	102 essays randomly chosen from the Argument Annotated Essays corpus	Krippendorff's agreement + Five-fold cross-validation
Ke et al. [76]	NLP+ML	Thesis strength+ Attribute scoring	Attribute features + Sentence features	ASAP corpus+CLC-FCE+Swedish corpus	QWK + Pearson's Correlation Coefficient (PC)
Jin et al. [22]	NLP+DL	TDNN: A Two-stage Deep Neural Network Model	Prompt-independent features+semantic, part-of-speech (POS) + Syntactic features	8 ASAP datasets	QWK + Mean square error (MSE)+ PCC + SCC
Amorim et al. [31]	NLP	N/A	Domain features+General features	1,840 essays of a standardised Brazilian national exam(high school)	Five-fold cross validation
Cummins and Rei [77]	NLP+ML	Neural Multi-task Learning	N/A	First Certificate in English (FCE) dataset	QWK + Google News embeddings + Spearman's rank correlation coefficient
Shin and Gierl [78]	NLP+ML	Deep-neural (or CNN)	Representative linguistic features	6 ASAP ++ datasets	QWK + Accuracy
Ikram and Castle [79]	NLP+ML	Coh-Metrix	Four new semantic features	Training: 1000 essays, testing: 786 essays	QWK + Adjacent Accuracy
Palma and Atkinson [80]	NLP	Discourse-based merges semantic and syntactic models	Readability features + Shallow linguistic features	8 ASAP datasets	QWK + Spearman correlation
Shin and Gierl [81]	NLP+ML+DL	Coh-Metrix + SVMs model + CNNs model	Deep/complex Language features + Coh-Metrix features	8 ASAP datasets	QWK
Yuan et al. [82]	DL	CNNs model	Linguistic and semantic features	Two Chinese essay datasets developed by themselves	Correlation Coefficient+ RMSE + Multiple Linear Regression (MLR)
Chen and Li [83]	NLP+DL	Hierarchical Recurrent Neural Networks	Sentence-level and Document-level +context information	8 ASAP datasets	Average QWK + QWK

TABLE 5. *(Continued.)* **Techniques used in existing AES models and their characteristics.**

Yang et al. [84]	NLP+DL	AGCE system+Bi-LSTM-CRF model	Hidden Markov Model (HMM)	1,000 pictures of Grade 3 students' essays	Kappa + Linear Weighted Kappa (LWK) + QWK
Ye and Manoharan [85]	NLP+DL	BERT model	Sentence features	Created a dataset with 400 samples	Pearson's correlation coefficient + Mean Squared Error (MSE)
Liang et al. [86]	NLP +DL	Siamese Bidirectional LSTM Neural Network	Inner-feature + cross-feature	8 ASAP datasets	QWK
Janda et al. [87]	NLP +Text Mining(TM)	Joint Effect model	Syntactic, sentiment + semantic features	8 ASAP datasets	QWK
Li et al. [88]	NLP+DL	Coherence-based Scoring with Self-Attention	Long-distance relationships across sentences	6 ASAP datasets +a new non-native speaker dataset	QWK
Devlin et al. [89]	NLP +DL	BERT: Pre-training of Deep Bidirectional Transformers	N/A	COLA corpus (10,000 sentences) +STS-B +Microsoft Research Paraphrase Corpus(MRPC)	F1-scores + MultiNLi Accuracy
Contreras et al. [90]	NLP + ML	Support Vector Machine (SVM) Model	Domain ontology+Numerical features+Parts of Speech count + Orthography + Similarity	Essays about Human Computer Interaction from real essay exams	Linear regression + Similarities
Farag and Yannakoudakis [91]	NLP+DL	Neural Single-Task Learning (STL)	Syntactic features	Wall Street Journal (WSJ)+Grammarly Corpus of Discourse Coherence (GCDC)	Three-way classification + Accuracy
Uto and Okano [92]	NLP+DL	DNN+IRT-based AES models	N-gram level features	ASAP datasets	Averaged score values + Standard Deviations (SD)
Wang et al. [93]	NLP+ML	Bi-directional +Dilated LSTM +Reinforcement Learning	Bag-of-word features + Content and style features + Linguistic features	ASAP datasets	QWK,Mean Square Error (MSE)
Jankowska et al. [94]	NLP	Common N-Gram (CNG)+linear SVM with SGD learning+NB	Bag-of-n-grams + N-gram features	3 ASAP datasets: set 2, set 7 and set 8	QWK
Do et al. [23]	NLP+DL	prompt + trait relation-aware cross-prompt trait scorer (ProTACT)	Topic-Coherence	ASAP and ASAP ++ datasets	Average QWK
Liu et al. [95]	NLP+graph neural networks (GNN)	GCN-based coherence model	latent +semantic feature	Grammarly Corpus of Discourse Coherence (GCDC) +TOEFL	Mean accuracy
Singh et al. [96]	NLP+ML	Hindi-AES model	6 classical features	Hindi Translated ASAP Corpus	QWK
Tashu [97]	DL	C-BGRU Siamese model	window-based feature	8 ASAP	Precision + Recall + F1-scores
Sun et al. [98]	NLP	Prompt Prediction and Matching model	prompt features	HSK + Dynamic Composition Corpus	QWK + PCC
Li et al. [99]	NLP+DL	Shared and Enhanced Deep Neural Network model	coherence /essay/prompt/relevance features	8 ASAP datasets	QWK
Mizumoto and Eguchi [14]	NLP+DL	AI language model	Linguistic features	TOEFL11	QWK Coefficients
Li et al. [100]	NLP+DL	DNN-AES model	deep semantic / multi-scale /linguistic / prompt-related features	8 ASAP datasets	Average QWK
Suresh et al. [101]	NLP+DL	AI based model	302 different features	8 ASAP datasets	MAE (Mean Absolute Error)
Gupta [102]	ML	Transformer models +data augmentation	N/A	8 ASAP datasets	k-learn Accuracy

TABLE 5. *(Continued.)* **Techniques used in existing AES models and their characteristics.**

Authors [Ref.]	Technique(s)	Model(s)	Feature(s)	Dataset(s)	Metrics
Altamimi [13]	NLP	GAI model	N/A	2 ASAP datasets(set 1 and set 8)	MAE (Mean Absolute Error)
Uto et al. [103]	Item Response Theory(IRT) + DL	IRT-based score-integration model	Mixed features from other model	8 ASAP datasets	average QWK + IRTscores

make up each sample. In addition, some researchers have studied the application of AES to different languages. Hirao et al. utilised a dataset named Goodwriting which consists of more than 800 essays authored by Japanese students studying abroad to create an AES system designed for Nonnative Japanese Learners [17].

2) EVALUATION METRICS

Various metrics have been applied to assess how well AES systems compare to a gold standard set by humans. These measures include correlation, accuracy, precision, recall, F1, and receiver Operating Characteristic (ROC). Among these measures, correlation is the most commonly used measure for evaluating the performance of AES systems. It measures the linear relationship between the expected and gold standard scores determined by humans and is calculated using the Pearson correlation coefficient. Yannakoudakis and Cummins suggested in their study that when assessing the efficacy of ATS systems, metrics of the agreement are preferable to measures of association (i.e., correlation) [15], as depicted:

$$C_\kappa = \frac{P_a - P_e(\kappa)}{1 - P_e(\kappa)} \tag{1}$$

Here, P_a is the actual percentage of agreement observed between the raters, while P_e is the percentage of agreement expected if the raters were just randomly guessing.

Accuracy, precision, and recall measure how accurately the AES system can classify text as belonging to a given score level. Accuracy is calculated below:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{2}$$

Whereas T denotes true and F denotes false, and P signifies positive while N denotes negative. Other metrics that may be used to assess how well AES systems are doing include F1 and ROC. The actual positive rate is shown against the false positive rate on a graph known as the ROC, whereas F1 measures the harmonic mean of accuracy and recall.

An evaluation method aims to assess the accuracy or performance of an AES model or system. The mapping of different evaluation metrics in AES Systems is demonstrated in Figure 5. Meanwhile, each of the evaluation matrices used in the screened papers is shown in Table 5. As it exhibits, the evaluation matrices used in the screened papers are not all identical, which makes it difficult to assess the goodness of the models, as the evaluation methods or criteria are not uniform. By looking at the data in the table, it is plain to see which matrices are widely used and accepted.

The Quadratic Weighted Kappa (QWK) agreement measure is a prominent metric widely utilized in various assessment contexts. It evaluates the degree of agreement among raters, particularly in scenarios where ordinal categorical ratings are employed. This metric is between 0 and 1, representing no agreement beyond chance and perfect agreement, respectively. However, it is imperative to note that a low consensus, compared to what would be anticipated by chance, could render QWK potentially misleading or even detrimental. To compute QWK, an N-by-N matrix of weights denoted as w is derived. This matrix encapsulates the differences between the scores assigned by raters. The weight calculation is expressed as follows:

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2} \tag{3}$$

Here, i and j represent the ratings provided by two distinct raters, while N denotes the total number of categories. The Quadratic Weighted Kappa[8] κ is determined through the following formula:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} \cdot O_{i,j}}{\sum_{i,j} w_{i,j} \cdot E_{i,j}} \tag{4}$$

Here, $O_{i,j}$ represents the observed frequency of agreement between raters for each combination of categories, while $E_{i,j}$ signifies the expected frequency of agreement if no correlation existed between raters.

The restriction of QWK in managing a high number of raters was discussed by Doewes et al. [18], who also proposed substitute metrics, like Fleiss kappa or Krippendorff's alpha, for situations requiring more than two raters. Another commonly used metric is the Mean Squared Error (MSE), which calculates the mean square of the mistakes or departures from the actual value. This metric can provide a good indication of how close the estimated values are to the valid values. Also, the Root Mean Squared Error (RMSE) is the square root of the MSE and is often used to compare different models. Additionally, since the Mean Absolute Error (MAE) calculates the average of the absolute errors

[8]Additional elucidation of the formula may be sought through reference to the corresponding website:https://www.kaggle.com/c/asap-aes. Consequently, to prevent redundancy, this article refrains from offering an extensive exposition thereof.

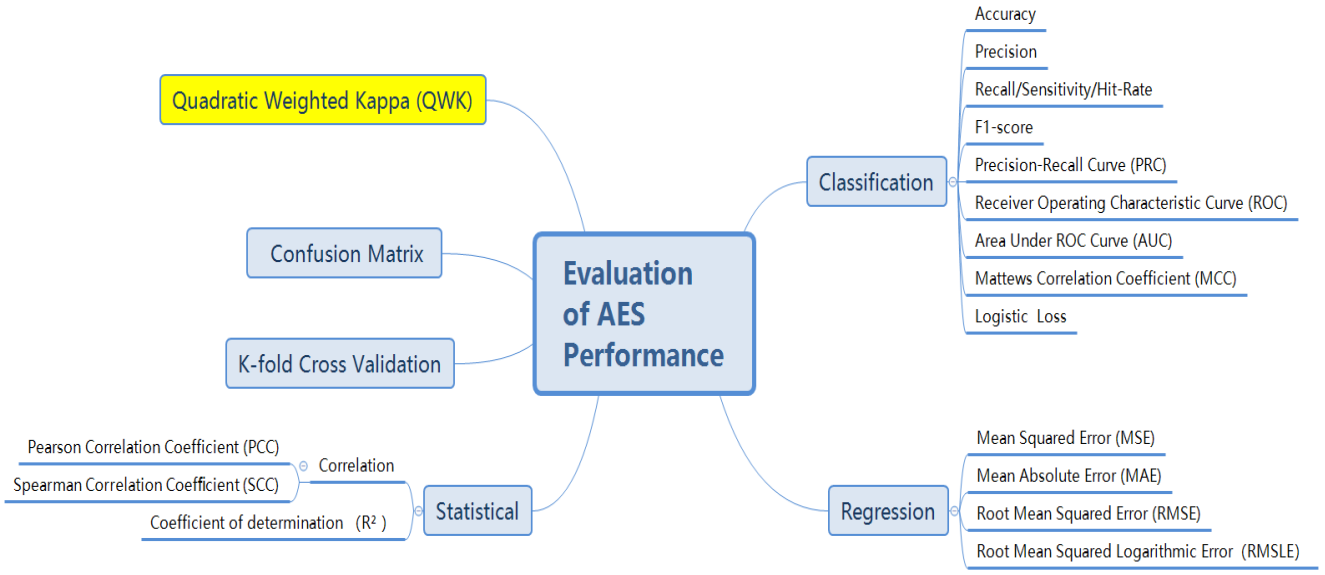

FIGURE 5. The mapping of different evaluation metrics in AES Systems.

or deviations from the genuine value, it is frequently used to assess how closely estimated values match valid values. Other often-used metrics include correlation measures like Pearson's correlation coefficient (PCC) and Spearman's correlation coefficient.

E. CHALLENGES OR LIMITATIONS OF THE CURRENT AES SYSTEMS (RQ4)

A description of the techniques and methods used in the existing AES models has been enumerated in Table 5. The explanations of these methods are required to give an insight into the functionality and characteristics of each AES model. Some authors of the literature papers provide the shortcomings of their models and suggest more study avenues. These limitations are the basis for any desired improvements. While pioneers have put much effort into developing and improving AES models, challenges still require pressing studies and investigation. These restrictions are derived from particular research.

In general, the limitations of existing AES models can be summed up in three aspects, which are listed as follows: 1. improving the performance of AES; 2. enriching the functions of AES; 3. AES assist tools.

1) IMPROVING THE PERFORMANCE OF AES MODELS

AES is developing technology as researchers have conducted extensive experiments and quantitative studies to enhance the precision and effectiveness of AES systems. For instance, the accuracy of their model's predicted scores, assessing different topics or styles of the essays, enabling the processing of more extended essays, and improving the robustness of AES models with some adversarial inputs.

Some progress has been made in recent years, as seen from the last section's benchmark, but it is still far from practical application. AES studies seem to have reached a plateau in improving the performance of different models or algorithms [19]. Fortunately, there is room for improvement, and many researchers provide suggestions for their work. Wang et al. suggest that soft multi-scale representation has advantages in processing long texts, and more progress may be made if language-specific data is added to the segment on a more reasonable magnitude [20]. In addition, there is a direction on how to boost the model's flexibility. The vast majority of the designed systems can only be adapted to assessing writing on a particular essay topic, which is challenging to meet the needs in real-life teaching scenarios. Kumar and Boulanger suggested assessing their system on essays composed in response to distinct prompts and exercising it on writings produced in response to a single set of source prompts to investigate how to broaden their approach to generalise their system [21]. Jin et al. specifically studied addressing the prompt-independent AES to improve the adaptability of their model [22]. They believe that the proposed TDNN might be enhanced by transferring non-target training data to the target prompt, and transfer learning methods are promising to prompt-independent AES. In order to successfully increase the accuracy of cross-prompt essay trait scoring, Do et al.concentrated on enhancing the trait scoring accuracy for essays written in response to various prompts [23]. To this end, they introduced a trait-similarity loss to encourage the model to learn similar representations for traits related to the prompts.

Besides considering the AES systems' accuracy, Automated Scoring system developers must consider the robustness of their systems. Hence, they are not susceptible to

any adversarial assault. Kabra et al. note that more efficient AES systems would be designed by combining proposed metrics to provide more holistic criteria for analysis and then strengthening the assessment suite with an emphasis on test type and student education level [24].

Another group of researchers studied models used to assess a specific aspect of the essay, such as coherence, articulation, and task response, to improve the model's overall characteristics. Mim et al.are determining how to use additional unannotated argumentation texts other than student essays for the suggested pre-training strategy [25]. There are different criteria for assessing whether a paper is good or bad. Many other aspects of these grading rules would be further explored, such as whether the paper is relevant to the topic.

2) ENRICH THE FUNCTIONS OF AES

Pioneers are not satisfied with enhancing the performance of models singly. Many researchers believe that they should enrich their functionality. They point out that it would be more meaningful if the models provided the author with writing feedback instead of merely assigning a score. Most of the research on AES is centred on evaluating score correctness or gives priority to developing scoring models to optimise agreement with human evaluators, which is still far from enough to put AES into the educational field to serve teachers and students [26]. Therefore, there are still significant Gaps in AES in this regard.

Kumar and Boulanger recommend written feedback to the author utilising characteristic scoring, such as identifying the source of a poor score for a particular characteristic [21]. Similarly, Hellman et al. believe that additional research is necessary to understand the route from subject localisation and holistic scoring to the most beneficial sorts of student feedback [27]. Different types of educational feedback should be analysed to see how they could be personalised. In addition to text, videos, peer interaction, practical examples, and other pedagogical interventions from the complete spectrum of possibilities could be given to enrich the types of feedback. Tsai et al. designed a system called LinggleWrite, which can assess essays and provide corrective feedback to improve users' writing abilities [28]. Their system has ample opportunity for improvement; for instance, introducing extra training data or generating false training data could be implemented. Reordering corrective suggestions so that the proposal most related to the original text is at the top is an intriguing direction to pursue. Detecting different kinds of mistakes at the granular level is an additional research direction.

Other researchers have extended AES research to language applications in their own countries. For example, for non-native Japanese learners, Hirao et al. developed an automated essay scoring (AES) system [17]; a similar model is UKARA 1.0 designed by eptiandri Septiandri and Winatmoko [29]. More and more researchers are working on the applications of AES in detecting non-English essay writing assessments, which will be a popular research trend in the future.

3) AES ASSIST TOOLS

Researchers or practitioners often need to experiment with various models to test their models' performance and compare their algorithms' strengths and weaknesses. Therefore, developing AES assist tools with good compatibility is very popular. Manabe and Hagiwara think that their designed ToolKit needs to include more features of the AES model and the methods used in the algorithms [30]. In addition, developing a corpus can also help advance AES. Amorim et al. created a new corpus of 1840 high-school students' essays from the Brazilian national exam and the subjectivity lexicons to assess how much rater bias impacts the efficiency of modern AES models [31].

V. DISCUSSION

A. WHAT SORT OF DATASET SHOULD BE COLLECTED?

Although there are currently a fraction of publicly available datasets, more is needed for studying the development of DL models and other algorithms. Collections of essays containing different stylistic categories, collections of model essays with professionally crafted annotations to give scores, and essays on different topics are encouraged to be collected.

In addition to papers in the dominant language, English, datasets in other languages are also worth collecting. However, this part of the work can be carried out when the mainstream language is more mature so that AES systems can benefit a larger group of people and play a more valuable role.

B. IS HUMAN-CENTRIC FEATURE EXTRACTION STILL NEEDED?

The necessity of human-centric feature extraction in AES remains a topic of ongoing deliberation and investigation. Despite the emergence of advanced algorithms in AI and ML capable of automated feature extraction from essays, specific circumstances exist where human-centric approaches retain significant importance.

In domains like AES, where interpretability, domain-specific expertise, and subjective human judgment hold primacy, human-centric feature extraction retains its indispensability. Human experts are adept at discerning subtle features that automated methods may overlook, contributing to a more nuanced understanding of essay content. Furthermore, feature extraction techniques are integral to the design of AES systems with feedback mechanisms, given their advantages in terms of interpretability and comprehensibility. However, it is worth noting that this approach often demands substantial technical effort to achieve high precision across a diverse range of articles.

Conversely, automated feature extraction methods may suffice in domains prioritizing large-scale data analysis and computational efficiency. For instance, Deep neural

networks (DNNs) employ a neural network technique that automates feature extraction, obviating the need for manual feature engineering. Successful deployment of neural models relies on extensive training data, particularly in languages other than English, where adequate data may be scarce. These methods exhibit adeptness in efficiently handling extensive datasets and unveiling patterns that may evade initial human observation.

To conclude, while the significance of human-centric feature extraction persists in specific contexts, the advent of automated methodologies has significantly reshaped the landscape of feature extraction. The selection between human-centric and automated approaches hinges upon the task's specific requirements, constraints, and objectives. Although several neural holistic scoring models have achieved notable advancements, integrating custom features via feature engineering holds the potential for further enhancement. Consequently, viewing feature-based and neural techniques as complementary rather than antagonistic frameworks is advisable.

C. WHAT IS THE LATEST TREND IN TERMS OF TECHNOLOGY IN AES?

As an artificial intelligence language model developed by OpenAI, ChatGPT-4 is not a specific technology or tool designed for AES tasks. However, the continued advancement and development of language models like ChatGPT-4 has the potential to impact AES in several ways.

First, the capacity of language models to produce well-written and cohesive text can potentially be used to generate automated essay responses. While current AES systems focus on scoring pre-written essays, future developments could allow for the generation of essays that meet specific criteria, such as prompt response and coherence, thus converting the task's original purpose of assessment into generation.

Second, pre-training and fine-tuning AES systems using massive computational language models such as ChatGPT-4 might enhance the precision and efficiency of AES systems. These models can pick up large amounts of textual data and capture linguistic nuances, improving their ability to score texts based on coherence, argumentation, and grammar.

Finally, the increasingly widespread use of language models in NLP can impact the field of AES by improving the accuracy and effectiveness of other NLP techniques related to AES, such as topic modelling and sentiment analysis. These models, in turn, can enhance the calibre of the training data and evaluate AES systems.

D. HOW SHOULD AES SYSTEMS BE ASSESSED?

When it comes to the current means of evaluating AES, more attention is paid in terms of the accuracy of the scores given by the systems. It seems skewed and biased to measure the system by the criterion of accuracy alone since it is well-known that grades(imbalanced datasets) obey a normal distribution. More kinds of metrics are encouraged in order to evaluate the various aspects of AES performance more comprehensively. However, this raises another problem, which is the inconsistent evaluation system. On the one hand, we encourage more researchers to use highly accepted and influential evaluation metrics like QWK to evaluate their systems; on the other hand, we also encourage other evaluation metrics to be widely accepted and recognised to improve the evaluation of AES systems.

In summary, there is no ''one-size-fits-all'' evaluation metric. When picking or specifying the appropriate evaluation metrics, it is essential to get to know the data and consider the AES system's objectives.

VI. LIMITATIONS

Although we have carefully designed inclusion and exclusion criteria and quality assessment checklists to filter the appropriate papers for our SLR, some limitations must be acknowledged. Firstly, we cannot ensure we have included all the relevant papers, as many digital databases like ACM digital library and ScienceDirect have yet to be searched. Another cause for worry is that because only English-published publications were considered in this evaluation, significant or pertinent research that was overlooked in journals with other language publications certainly exists. A final shortcoming is that the literature we have focused on does not span enough time, with the earliest being 2018 and the latest being the end of 2023. We started retrieving articles from 2018 because DL has been brought into the spotlight since 2018, and various new AES models have sprung up. While this gives us an idea of the most popular and cutting-edge AES technology in recent years, more is needed to investigate its development and processes as previous excellent technology is ignored.

VII. CONCLUSION

This study conducted a systematic literature review to assess the competence of Automated Essay Scoring (AES) systems in real-life education scenarios over the past six years. A comprehensive examination of relevant literature served as the research methodology, in which the study formulated several significant research questions to identify and evaluate current AES algorithms, feature extraction methods, metric evaluations and limitations. The primary objective was to discern potential areas for improvement or enhancement by analyzing recent studies on AES approaches. In order to ensure the accuracy and applicability of the investigations, a review process and specific quality evaluation criteria were developed. These criteria delineated steps such as research identification, paper selection, evaluation of paper quality, and data extraction and synchronization.

Although encountered, validity concerns were predominantly identified early in the study and appropriately addressed. Pertinent primary studies were conducted using these methods, and the quality of this study was assessed. Data extraction and synchronization were conducted using information retrieved from the initial research, with data sourced from four significant internet database sources.

Consequently, the research questions were addressed, and the research objectives were deemed fulfilled. Studies aimed at improving current AES approaches were identified and summarized.

In summary, the findings underscore the ongoing necessity for advancements in AES systems. Despite the existence of diverse models, further improvements are imperative to address various constraints identified in this study. These include enhancing scalability to accommodate diverse essay topics and styles, refining score prediction accuracy, and fortifying outcome reliability. Additionally, bolstering the resilience of AES models against adversarial inputs, enriching system functionality, and developing tailored assistive tools are vital considerations for future research endeavors in AES methodologies. Consequently, while existing AES systems exhibit commendable accuracy within specific contexts, they still need to fully meet the demands of educators and students in authentic teaching scenarios. Enhancements must focus not only on scalability and accuracy but also on robustness, functionality, and provision of assistive tools to advance AES systems in practical educational settings.

REFERENCES

[1] M. Liu, "Research on the key technology of the automatic scoring of the college entrance examination essay," Harbin Inst. Technol., Harbin, China, Tech. Rep., pp. 173–181, 2015, vol. 30, no. 6.

[2] T. Dasgupta, A. Naskar, L. Dey, and R. Saha, "Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring," in *Proc. 5th Workshop Natural Lang. Process. Techn. Educ. Appl.*, 2018, pp. 93–102.

[3] T. K. Landauer, "Automatic essay assessment," *Assessment Educ., Princ., Policy Pract.*, vol. 10, no. 3, pp. 295–308, Nov. 2003.

[4] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: A systematic literature review," *Artif. Intell. Rev.*, vol. 55, no. 3, pp. 2495–2527, Mar. 2022.

[5] S. Keele, "Guidelines for performing systematic literature reviews in software engineering," Keele Univ., Durham Univ., U.K., Joint Rep. EBSE 2007-001, 2007.

[6] M. A. Hussein, H. Hassan, and M. Nassef, "Automated language essay scoring systems: A literature review," *PeerJ Comput. Sci.*, vol. 5, p. e208, Aug. 2019.

[7] R. H. Chassab, L. Q. Zakaria, and S. Tiun, "Automatic essay scoring: A review on the feature analysis techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 10, pp. 1–13, 2021.

[8] C. T. Lim, C. H. Bong, W. S. Wong, and N. K. Lee, "A comprehensive review of automated essay scoring (AES) research and development," *Pertanika J. Sci. Technol.*, vol. 29, no. 3, pp. 1875–1899, Jul. 2021.

[9] J. G. Borade and L. D. Netak, "Automated grading of essays: A review," in *Proc. Int. Conf. Intell. Hum. Comput. Interact.*, Daegu, South Korea. Cham, Switzerland: Springer, 2021, pp. 238–249.

[10] J. Zhang and J. Liu, "Deep-neural automated essay scoring: A review," in *Proc. 3rd Int. Conf. Comput. Inf. Big Data Appl.*, Mar. 2022, pp. 1–4.

[11] M. Uto, "A review of deep-neural automated essay scoring models," *Behaviormetrika*, vol. 48, no. 2, pp. 459–484, Jul. 2021.

[12] D. Ifenthaler, "Automated essay scoring systems," in *Handbook of Open, Distance and Digital Education*. Singapore: Springer, 2023, pp. 1057–1071.

[13] A. B. Altamimi, "Effectiveness of ChatGPT in essay autograding," in *Proc. Int. Conf. Comput., Electron. Commun. Eng. (iCCECE)*, Aug. 2023, pp. 102–106.

[14] A. Mizumoto and M. Eguchi, "Exploring the potential of using an AI language model for automated essay scoring," *Res. Methods Appl. Linguistics*, vol. 2, no. 2, Aug. 2023, Art. no. 100050.

[15] H. Yannakoudakis and R. Cummins, "Evaluating the performance of automated text scoring systems," in *Proc. 10th Workshop Innov. Use NLP Building Educ. Appl.*, 2015, pp. 213–223.

[16] X. Ye and S. Manoharan, "Performance comparison of automated essay graders based on various language models," in *Proc. IEEE Int. Conf. Comput. (ICOCO)*, Nov. 2021, pp. 152–157.

[17] R. Hirao, M. Arai, H. Shimanaka, S. Katsumata, and M. Komachi, "Automated essay scoring system for nonnative Japanese learners," in *Proc. 12th Lang. Resour. Eval. Conf.*, May 2020, pp. 1250–1257.

[18] A. Doewes, N. Kurdhi, and A. Saxena, "Evaluating quadratic weighted Kappa as the standard performance metric for automated essay scoring," in *Proc. 16th Int. Conf. Educ. Data Mining*, Jul. 2023, pp. 103–113.

[19] P. Uria Rodriguez, A. Jafari, and C. M. Ormerod, "Language models and automated essay scoring," 2019, *arXiv:1909.09482*.

[20] Y. Wang, C. Wang, R. Li, and H. Lin, "On the use of BERT for automated essay scoring: Joint learning of multi-scale essay representation," 2022, *arXiv:2205.03835*.

[21] V. S. Kumar and D. Boulanger, "Automated essay scoring and the deep learning black box: How are rubric scores determined?" *Int. J. Artif. Intell. Educ.*, vol. 31, no. 3, pp. 538–584, Sep. 2021.

[22] C. Jin, B. He, K. Hui, and L. Sun, "TDNN: A two-stage deep neural network for prompt-independent automated essay scoring," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1088–1097.

[23] H. Do, Y. Kim, and G. G. Lee, "Prompt- and trait relation-aware cross-prompt essay trait scoring," 2023, *arXiv:2305.16826*.

[24] A. Kabra, M. Bhatia, Y. K. Singla, J. Jessy Li, and R. R. Shah, "Evaluation toolkit for robustness testing of automatic essay scoring systems," in *Proc. 5th Joint Int. Conf. Data Sci. Manage. Data*, Jan. 2022, pp. 90–99.

[25] F. S. Mim, N. Inoue, P. Reisert, H. Ouchi, and K. Inui, "Corruption is not all bad: Incorporating discourse structure into pre-training via corruption for essay scoring," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2202–2215, 2021.

[26] M. Hardy, "Toward educator-focused automated scoring systems for reading and writing," 2021, *arXiv:2112.11973*.

[27] S. Hellman, W. Murray, A. Wiemerslage, M. Rosenstein, P. Foltz, L. Becker, and M. Derr, "Multiple instance learning for content feedback localization without annotation," in *Proc. 15th Workshop Innov. Use NLP Building Educ. Appl.*, 2020, pp. 30–40.

[28] C.-T. Tsai, J.-J. Chen, C.-Y. Yang, and J. S. Chang, "LinggleWrite: A coaching system for essay writing," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstration*, 2020, pp. 127–133.

[29] A. Akbar Septiandri and Y. Ardhito Winatmoko, "UKARA 1.0 challenge track 1: Automatic short-answer scoring in bahasa Indonesia," 2020, *arXiv:2002.12540*.

[30] H. Manabe and M. Hagiwara, "EXPATS: A toolkit for explainable automated text scoring," 2021, *arXiv:2104.03364*.

[31] E. Amorim, M. Cançado, and A. Veloso, "Automated essay scoring in the presence of biased ratings," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 229–237.

[32] Y. Gao, R. Yu, and X. Duan, "An English handwriting evaluation algorithm based on CNNs," in *Proc. Int. Conf. Artif. Intell. Adv. Manuf. (AIAM)*, Oct. 2019, pp. 18–21.

[33] F. Chen, "Automatic integrated scoring model for English composition oriented to part-of-speech tagging," *Complexity*, vol. 2021, pp. 1–13, May 2021.

[34] J. Zhao and J. Chen, "Automatic scoring system for CET-4 compositions based on Seq2Seq+Bi-LSTM model," in *Proc. IEEE 3rd Int. Conf. Autom., Electron. Electr. Eng. (AUTEEE)*, Nov. 2020, pp. 333–336.

[35] M. Hsiao and M. Hung, "Construction of an artificial intelligence writing model for English based on fusion neural network model," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–12, May 2022.

[36] Z. He, "English grammar error detection using recurrent neural networks," *Sci. Program.*, vol. 2021, pp. 1–8, Jul. 2021.

[37] R. Xiao, W. Guo, Y. Zhang, X. Ma, and J. Jiang, "Machine learning-based automated essay scoring system for Chinese proficiency test (HSK)," in *Proc. 4th Int. Conf. Natural Lang. Process. Inf. Retr.*, Dec. 2020, pp. 18–23.

[38] T. Elks, "Using transfer learning to automatically mark l2 writing texts," in *Proc. Student Res. Workshop Associated RANLP*, 2021, pp. 51–57.

[39] M. Beseiso, O. A. Alzubi, and H. Rashaideh, "A novel automated essay scoring approach for reliable higher educational assessments," *J. Comput. Higher Educ.*, vol. 33, no. 3, pp. 727–746, Dec. 2021.

[40] H. Chimingyang, "An automatic system for essay questions scoring based on LSTM and word embedding," in *Proc. 5th Int. Conf. Inf. Sci., Comput. Technol. Transp. (ISCTT)*, Nov. 2020, pp. 355–364.

[41] A. N. Oktaviani, M. Z. Alief, L. Santiar, P. D. Purnamasari, and A. A. P. Ratna, "Automatic essay grading system for Japanese language exam using CNN-LSTM," in *Proc. 17th Int. Conf. Quality Res. (QIR), Int. Symp. Electr. Comput. Eng.*, Oct. 2021, pp. 164–169.

[42] H. M. Balaha and M. M. Saafan, "Automatic exam correction framework (AECF) for the MCQs, essays, and equations matching," *IEEE Access*, vol. 9, pp. 32368–32389, 2021.

[43] L. B. Das, C. V. Raghu, G. Jagadanand, R. A. R. George, P. Yashasawi, N. A. A. Kumaran, and V. K. Patnaik, "FACToGRADE: Automated essay scoring system," in *Proc. IEEE Int. Conf. Ind. 4.0, Artif. Intell., Commun. Technol. (IAICT)*, Jul. 2022, pp. 42–48.

[44] Z. Wang, J. Liu, and R. Dong, "Intelligent auto-grading system," in *Proc. 5th IEEE Int. Conf. Cloud Comput. Intell. Syst. (CCIS)*, Nov. 2018, pp. 430–435.

[45] Y. Cong, "Intelligent English writing automatic scoring system based on data fusion algorithm," in *Proc. IEEE Asia–Pacific Conf. Image Process., Electron. Comput. (IPEC)*, Apr. 2022, pp. 1079–1082.

[46] A. A. P. Ratna, L. Santiar, I. Ibrahim, P. D. Purnamasari, D. L. Luhurkinanti, and A. Larasati, "Latent semantic analysis and winnowing algorithm based automatic Japanese short essay answer grading system comparative performance," in *Proc. IEEE 10th Int. Conf. Awareness Sci. Technol. (iCAST)*, Oct. 2019, pp. 1–7.

[47] Z. Chen and Y. Zhou, "Research on automatic essay scoring of composition based on CNN and OR," in *Proc. 2nd Int. Conf. Artif. Intell. Big Data (ICAIBD)*, May 2019, pp. 13–18.

[48] W. Wei, L. Yong-An, M. Lun, and Q. Qianqian, "Research on error detection technology of English writing based on recurrent neural network," in *Proc. Int. Conf. Big Data Anal. Comput. Sci. (BDACS)*, Jun. 2021, pp. 209–214.

[49] S. Vajjala, "Automated assessment of non-native learner essays: Investigating the role of linguistic features," *Int. J. Artif. Intell. Educ.*, vol. 28, no. 1, pp. 79–105, Mar. 2018.

[50] X. Li, H. Yang, S. Hu, J. Geng, K. Lin, and Y. Li, "Enhanced hybrid neural network for automated essay scoring," *Expert Syst.*, vol. 39, no. 10, Dec. 2022, Art. no. e13068.

[51] M. Beseiso and S. Alzahrani, "An empirical analysis of BERT embedding for automated essay scoring," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 10, pp. 1–15, 2020.

[52] B. Nicula, M. Dascalu, N. N. Newton, E. Orcutt, and D. S. McNamara, "Automated paraphrase quality assessment using language models and transfer learning," *Computers*, vol. 10, no. 12, p. 166, Dec. 2021.

[53] J. Hoblos, "Experimenting with latent semantic analysis and latent Dirichlet allocation on automated essay grading," in *Proc. 7th Int. Conf. Social Netw. Anal., Manage. Secur. (SNAMS)*, Dec. 2020, pp. 1–7.

[54] A. Sharma and D. B. Jayagopi, "Handwritten essay grading on mobiles using MDLSTM model and word embeddings," in *Proc. 11th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2018, pp. 1–8.

[55] V. Kumar and D. Boulanger, "Explainable automated essay scoring: Deep learning really has pedagogical value," *Frontiers Educ.*, vol. 5, Oct. 2020, Art. no. 572367.

[56] S. Jeon and M. Strube, "Countering the influence of essay length in neural essay scoring," in *Proc. 2nd Workshop Simple Efficient Natural Lang. Process.*, 2021, pp. 32–38.

[57] M. Cozma, A. M. Butnaru, and R. Tudor Ionescu, "Automated essay scoring with string kernels and word embeddings," 2018, *arXiv:1804.07954*.

[58] R. Kumar, S. Mathias, S. Saha, and P. Bhattacharyya, "Many hands make light work: Using essay traits to automatically score essays," 2021, *arXiv:2102.00781*.

[59] J. Liu, Y. Xu, and Y. Zhu, "Automated essay scoring based on two-stage learning," 2019, *arXiv:1901.07744*.

[60] Y. Farag, H. Yannakoudakis, and T. Briscoe, "Neural automated essay scoring and coherence modeling for adversarially crafted input," 2018, *arXiv:1804.06898*.

[61] H. Zhang and D. Litman, "Co-attention based neural network for source-dependent essay scoring," 2019, *arXiv:1908.01993*.

[62] S. Ludwig, C. Mayer, C. Hansen, K. Eilers, and S. Brandt, "Automated essay scoring using transformer models," *Psych*, vol. 3, no. 4, pp. 897–915, 2021.

[63] Y.-J. Jong, Y.-J. Kim, and O.-C. Ri, "Improving performance of automated essay scoring by using back-translation essays and adjusted scores," *Math. Problems Eng.*, vol. 2022, pp. 1–10, Jun. 2022.

[64] C. Ormerod, A. Jafari, S. Lottridge, M. Patel, A. Harris, and P. van Wamelen, "The effects of data size on automated essay scoring engines," 2021, *arXiv:2108.13275*.

[65] C. M. Ormerod, A. Malhotra, and A. Jafari, "Automated essay scoring using efficient transformer-based language models," 2021, *arXiv:2102.13136*.

[66] P. Muangkammuen and F. Fukumoto, "Multi-task learning for automated essay scoring with sentiment analysis," in *Proc. 1st Conf. Asia–Pacific Chapter Assoc. Comput. Linguistics 10th Int. Joint Conf. Natural Lang. Process., Student Res. Workshop*, 2020, pp. 116–123.

[67] M. Uto, Y. Xie, and M. Ueno, "Neural automated essay scoring incorporating handcrafted features," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 6077–6088.

[68] R. Yang, J. Cao, Z. Wen, Y. Wu, and X. He, "Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2020, pp. 1560–1569.

[69] R. Ridley, L. He, X. Dai, S. Huang, and J. Chen, "Prompt agnostic essay scorer: A domain generalization approach to cross-prompt automated essay scoring," 2020, *arXiv:2008.01441*.

[70] H. Zhang and D. Litman, "Automated topical component extraction using neural network attention scores from source-based essay scoring," 2020, *arXiv:2008.01809*.

[71] E. Mayfield and A. W. Black, "Should you fine-tune BERT for automated essay scoring?" in *Proc. 15th Workshop Innov. Use NLP Building Educ. Appl.*, 2020, pp. 151–162.

[72] C. L. M. Jeronimo, C. E. C. Campelo, L. B. Marinho, A. Sales, A. Veloso, and R. Viola, "Computing with subjectivity lexicons," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 3272–3280.

[73] S. J. Berggren, T. Rama, and L. Øvrelid, "Regression or classification? Automated essay scoring for Norwegian," in *Proc. 14th Workshop Innov. Use NLP Building Educ. Appl.*, 2019, pp. 92–102.

[74] F. Nadeem, H. Nguyen, Y. Liu, and M. Ostendorf, "Automated essay scoring with discourse-aware neural models," in *Proc. 14th Workshop Innov. Use NLP Building Educ. Appl.*, 2019, pp. 484–493.

[75] W. Carlile, N. Gurrapadi, Z. Ke, and V. Ng, "Give me more feedback: Annotating argument persuasiveness and related attributes in student essays," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 621–631.

[76] Z. Ke, H. Inamdar, H. Lin, and V. Ng, "Give me more feedback II: Annotating thesis strength and related attributes in student essays," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 3994–4004.

[77] R. Cummins and M. Rei, "Neural multi-task learning in automated assessment," 2018, *arXiv:1801.06830*.

[78] J. Shin and M. J. Gierl, "Evaluating coherence in writing: Comparing the capacity of automated essay scoring technologies," *J. Appl. Test. Technol.*, vol. 10, pp. 4–20, Aug. 2022.

[79] A. Ikram and B. Castle, "Automated essay scoring (AES); a semantic analysis inspired machine learning approach: An automated essay scoring system using semantic analysis and machine learning is presented in this research," in *Proc. 12th Int. Conf. Educ. Technol. Comput.*, Oct. 2020, pp. 147–151.

[80] D. Palma and J. Atkinson, "Coherence-based automatic essay assessment," *IEEE Intell. Syst.*, vol. 33, no. 5, pp. 26–36, Sep. 2018.

[81] J. Shin and M. J. Gierl, "More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms," *Lang. Test.*, vol. 38, no. 2, pp. 247–272, Apr. 2021.

[82] S. Yuan, T. He, H. Huang, R. Hou, and M. Wang, "Automated Chinese essay scoring based on deep learning," *Comput., Mater. Continua*, vol. 65, no. 1, pp. 817–833, 2020.

[83] M. Chen and X. Li, "Relevance-based automated essay scoring via hierarchical recurrent model," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Nov. 2018, pp. 378–383.

[84] Y. Yang, L. Xia, and Q. Zhao, "An automated grader for Chinese essay combining shallow and deep semantic attributes," *IEEE Access*, vol. 7, pp. 176306–176316, 2019.

[85] X. Ye and S. Manoharan, "Marking essays automatically," in *Proc. 4th Int. Conf. E-Educ., E-Bus. E-Technol.*, Jun. 2020, pp. 56–60.

[86] G. Liang, B.-W. On, D. Jeong, H.-C. Kim, and G. Choi, "Automated essay scoring: A Siamese bidirectional LSTM neural network architecture," *Symmetry*, vol. 10, no. 12, p. 682, Dec. 2018.

[87] H. K. Janda, A. Pawar, S. Du, and V. Mago, "Syntactic, semantic and sentiment analysis: The joint effect on automated essay evaluation," *IEEE Access*, vol. 7, pp. 108486–108503, 2019.
[88] X. Li, M. Chen, J. Nie, Z. Liu, Z. Feng, and Y. Cai, "Coherence-based automated essay scoring using self-attention," in *Proc. 17th China Nat. Conf. CCL*, Changsha, China. Cham, switzerland: Springer, 2018, pp. 386–397.
[89] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[90] J. O. Contreras, S. Hilles, and Z. B. Abubakar, "Automated essay scoring with ontology based on text mining and NLTK tools," in *Proc. Int. Conf. Smart Comput. Electron. Enterprise (ICSCEE)*, Jul. 2018, pp. 1–6.
[91] Y. Farag and H. Yannakoudakis, "Multi-task learning for coherence modeling," 2019, *arXiv:1907.02427*.
[92] M. Uto and M. Okano, "Robust neural automated essay scoring using item response theory," in *21st Int. Conf. Artif. Intell.* Cham, Switzerland: Springer, 2020, pp. 549–561.
[93] Y. Wang, Z. Wei, Y. Zhou, and X. Huang, "Automatic essay scoring incorporating rating schema via reinforcement learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 791–797.
[94] M. Jankowska, C. Conrad, J. Harris, and V. Kešelj, "N-gram based approach for automatic prediction of essay rubric marks," in *Proc. Can. Conf. Artif. Intell.*, Toronto, ON, Canada. Cham, switzerland: Springer, 2018, pp. 298–303.
[95] W. Liu, X. Fu, and M. Strube, "Modeling structural similarities between documents for coherence assessment with graph convolutional networks," 2023, *arXiv:2306.06472*.
[96] S. Singh, A. Pupneja, S. Mital, C. Shah, M. Bawkar, L. P. Gupta, A. Kumar, Y. Kumar, R. Gupta, and R. Ratn Shah, "H-AES: Towards automated essay scoring for Hindi," 2023, *arXiv:2302.14635*.
[97] T. M. Tashu, "Off-topic essay detection using C-BGRU Siamese," in *Proc. IEEE 14th Int. Conf. Semantic Comput. (ICSC)*, Feb. 2020, pp. 221–225.
[98] J. Sun, T. Song, J. Song, and W. Peng, "Improving automated essay scoring by prompt prediction and matching," *Entropy*, vol. 24, no. 9, p. 1206, Aug. 2022.
[99] X. Li, M. Chen, and J.-Y. Nie, "SEDNN: Shared and enhanced deep neural network model for cross-prompt automated essay scoring," *Knowl.-Based Syst.*, vol. 210, Dec. 2020, Art. no. 106491.
[100] F. Li, X. Xi, Z. Cui, D. Li, and W. Zeng, "Automatic essay scoring method based on multi-scale features," *Appl. Sci.*, vol. 13, no. 11, p. 6775, Jun. 2023.
[101] V. Suresh, R. Agasthiya, J. Ajay, A. A. Gold, and D. Chandru, "AI based automated essay grading system using NLP," in *Proc. 7th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, May 2023, pp. 547–552.
[102] K. Gupta, "Data augmentation for automated essay scoring using transformer models," in *Proc. Int. Conf. Artif. Intell. Smart Commun. (AISC)*, Jan. 2023, pp. 853–857.
[103] M. Uto, I. Aomi, E. Tsutsumi, and M. Ueno, "Integration of prediction scores from various automated essay scoring models using item response theory," *IEEE Trans. Learn. Technol.*, vol. 16, no. 6, pp. 983–1000, Dec. 2023.

WENBO XU received the B.S. and M.S. degrees in electronics and information engineering from Guilin University of Electronic Science and Technology, China, in 2016. He is currently pursuing the Ph.D. degree in artificial intelligence with the Faculty of Computer Science and Information Technology, University of Malaya (UM), Malaysia. His research interests include machine learning, deep learning, artificial intelligence in education, and natural language processing.

ROHANA MAHMUD received the Ph.D. degree from The University of Manchester, U.K.

She is currently an Associate Professor with the Department of Artificial Intelligence (AI), Faculty of Computer Science and Information Technology, University of Malaya (UM), Malaysia. She has been teaching various courses in AI and computer sciences disciplines. As her expertise is in *Natural Language Processing* (NLP), *Expert Systems*, *Machine Learning*, and Principles of Data Science, hence, she had carried out a few collaborative research with experts from different fields, such as *Language and Linguistics*. She was an Associate Member of the Student Soft Skills (SERU) Center and was one of the members of the University Women Association (UWA) Committee. She also gained the Master's Certificate in Leadership after successfully completing a Women's Leadership Course for the Institutes of Higher Learning from Tun Fatimah Hashim's Women Leadership Centre, Universiti Kebangsaan Malaysia.

WAI LAM HOO (Member, IEEE) received the B.Sc. degree (Hons.) in computer science, majoring in artificial intelligence and the Ph.D. degree from Universiti Malaya (UM), Kuala Lumpur, Malaysia, in 2010 and 2015, respectively.

He is currently a Senior Lecturer in UM. His research interest includes computer vision, image/video processing, and scene understanding.

Dr. Hoo is currently the chair for IEEE Computer Intelligence Society Malaysia Chapter. He was the local arrangement chair for IEEE International Conference of Image Processing in 2023, The 13th International Conference on Intelligent Robotics and Applications in 2020, IEEE 21st International Workshop on Multimedia Signal Processing in 2019 and The 3rd IAPR Asian Conference on Pattern Recognition in 2015. He was the publication chair for Asia-Pacific Signal and Information Processing Association Annual Submit and Conference in 2017.