Optimizing Automated Essay Scoring: A Comparative Study of Machine Learning Approaches with a Focus on Ensemble Methods

Kornwipa Poonpon[1], Wirapong Chansanam[1]

[1] Khon Kaen University, Thailand

Correspondence: Wirapong Chansanam, Khon Kaen University, Thailand.

Received: October 30, 2024 Accepted: February 24, 2025 Online Published: April 17, 2025

doi:10.5430/wjel.v15n5p272 URL: https://doi.org/10.5430/wjel.v15n5p272

Abstract

This study examines the optimization of Automated Essay Scoring (AES) systems for English language writing using advanced machine learning techniques, focusing on ensemble methods to enhance accuracy, consistency, and interpretability. An English written corpus includes a total of 17,793 English essays: 12,976 from the Automated Student Assessment Prize (ASAP) dataset and 4,817 from the Khon Kaen University Academic English Language Test (KKU-AELT). Linguistic features and semantic content critical to English writing proficiency were assessed using BERT, XGBoost, and Neural Networks models. Combining these models with Ridge Regression, the ensemble approach substantially reduced Root Mean Squared Error (RMSE) while balancing Cohen's Kappa and Quadratic Weighted Kappa scores, highlighting interpretive alignment challenges. The SHAP values were employed for feature importance analysis, and Bayesian optimization was applied for hyperparameter tuning, enhancing model transparency. The findings highlight the potential of ensemble AES to evaluate diverse aspects of English such as argumentation, coherence, and vocabulary complexity—applicable to various domains, from applied linguistics to literature and translation studies. The research offers scalable solutions for teaching and assessment, aligning AES systems with the pedagogical goals of supporting skill acquisition and providing actionable feedback. The study concludes that advanced AES models can serve as valuable complementary tools in language assessment, assisting teachers by providing consistent, detailed insights that foster English writing proficiency and skills development across diverse educational contexts.

Keywords: Automated Essay Scoring (AES), English writing, second language assessment, non-native speakers, machine learning, ensemble methods, model interpretability

1. Introduction

Essay writing is fundamental to English language learning, particularly for English as a Foreign Language (EFL) learners. Developing the ability to express thoughts, ideas, and arguments coherently and structured is crucial for effective communication in academic and professional settings (Oshima & Hogue, 2007; Weigle, 2002). For EFL learners, mastering essay writing skills not only enhances their linguistic proficiency but also fosters critical thinking, creativity, and analytical reasoning (Shermis & Wilson, 2024; Tracy-Ventura & Paquot, 2020). As a key assessment tool, essay writing provides a comprehensive view of learners' language abilities, including language knowledge (e.g., grammar, vocabulary, discourse skills) and strategic competence (e.g., goal setting, planning and control of execution) (Bachman & Palmer, 1996; Hyland, 2003). Moreover, writing essays helps learners refine their language skills through practice and feedback, which are essential for continuous improvement (Guo, 2023; Manipatruni et al., 2024). Therefore, supporting EFL learners in developing their essay writing skills is vital for their overall language learning journey and academic success.

Scoring essay writing is critical to assessing language proficiency, particularly in EFL contexts. Validity, reliability, and practicality are key factors to ensure that scoring systems effectively evaluate writing skills (Green, 2022; Pack et al., 2024). Traditionally, human raters have scored essays, relying on rubrics that assess various aspects, such as topic development, paragraph organization, coherence, and linguistic accuracy (e.g., Zribi & Smaoui, 2021). However, human scoring can be time-consuming and subject to variability, as raters may interpret criteria differently or experience fatigue over time (Toranj & Ansari, 2012).

In contrast, Automated Essay Scoring (AES) systems offer a promising alternative by providing consistent and reliable scores. AES systems use machine learning algorithms to evaluate essays based on predefined criteria, similar to human raters (Ramalingam et al., 2018). These systems can process large volumes of essays quickly and accurately, reducing the burden on human raters and minimizing scoring inconsistencies (Toranj & Ansari, 2012). While AES systems have shown strong correlations with human scores, they also provide detailed feedback on specific essay traits, such as content relevance, coherence, and linguistic range, which can be invaluable for students seeking to improve their writing skills (Manipatruni et al., 2024). Despite these advantages, AES systems still face challenges in fully replicating the nuanced judgments of human raters, particularly in assessing complex aspects of writing like creativity and argumentation. Nonetheless, integrating AES with human scoring can enhance the sustainability of assessment processes by providing efficient, consistent, and reliable evaluations that support both teaching and learning outcomes (Toranj & Ansari, 2012; Wei et al., 2023).

2. Literature Review

Second language writing assessment

Second language writing assessment refers to the evaluation of writing proficiency in a language that is not the learner's first language (Weigle, 2002). The primary goal of assessing writing is to make inferences about learners' language abilities (Swales & Feak, 2012). This field encompasses various methodologies and technologies aimed at understanding and improving L2 writing performance, which is influenced by factors, mainly fluency, accuracy, and complexity (Hamp-Lyons & Kroll, 1997; Hyland, 2003; Swales & Feak, 2012; Zhang, 2025). To guide raters in evaluating specific aspects of writing, scoring rubrics are employed, incorporating various criteria that reflect these writing qualities. For example, common essay writing criteria include cohesion and coherence, structure and organization, and language use (Cumming et al., 2005; Hyland, 2003; Srisawat & Poonpon, 2023). Therefore, clearly defined scoring criteria are crucial for raters or teachers to ensure consistency and reliability in their evaluations.

Essay scoring traditionally relies on human raters, who possess several key strengths. First, human raters are adept at (a) cognitively processing the information presented in a text, (b) integrating it with their existing knowledge, and (c) forming a judgment on the text's quality based on their comprehensive understanding. Trained human raters are particularly skilled at recognizing and appreciating a writer's creative and stylistic elements (e.g., artistic, ironic, rhetorical devices). Additionally, they can assess how well an essay's content matches the prompt. Human raters can also judge an examinee's critical thinking skills, including the quality of the argumentation and the factual accuracy of the claims presented in the essay. This nuanced evaluation is essential for providing detailed feedback that supports learning and improvement. While valuable, human scoring has its limitations. Recruiting and training qualified raters are challenging, requiring extensive monitoring and potential retraining to ensure consistent, high-quality assessments (Zhang, 2013). Furthermore, human raters are susceptible to errors and biases, including varying degrees of severity or leniency, inconsistency, halo effects, and differing scoring perceptions (Bejar, 2011).

Automated Essay Scoring (AES)

Recent advancements have led to the widespread adoption of automated writing evaluation (AWE) systems in language assessment, including Automated Essay Scoring (AES). These technologies enhance accuracy, provide feedback, and offer efficient, consistent, and scalable evaluation when used in conjunction with human raters (Geçkin et al., 2023; Klebanov & Madnani, 2022; Zribi & Smaoui, 2021). As educational institutions increasingly adopt digital learning platforms and face growing student populations, reliable and sophisticated AES systems have become more pressing than ever. The application of AES is increasingly significant in educational technology. Xiao et al. (2024) demonstrate the effectiveness of Large Language Models (LLMs), such as GPT-4 and fine-tuned GPT-3.5, in AES. Their experiments show that these models provide superior accuracy, consistency, and generalizability compared to traditional grading methods.

Additionally, LLMs enhance human graders' performance, allowing novices to match expert accuracy and experts to improve efficiency. Lahitani et al. (2016) examine the use of term frequency-inverse document frequency (TF-IDF) and cosine similarity to rank documents based on similarity, emphasizing computational methods in text evaluation. Yang et al. (2022) explore integrating AES in teaching to enhance Chinese EFL learners' assessment literacy. Their findings suggest that suitable AES integration and guidance help learners critically assess scores and effectively utilize feedback, underscoring AES tools' potential in advancing education. This research aims to advance the field of AES by exploring and optimizing machine learning approaches, with a particular focus on ensemble methods.

The assessment of essays presents unique challenges in natural language processing and machine learning. Unlike objective question formats, essays require evaluating complex linguistic features, coherence, argumentation, and adherence to prompt specifications. Traditional rubric-based manual scoring, while valuable, is time-consuming and can be subject to inter-rater variability. The use of machine learning and natural language processing (NLP) techniques to improve AES and placement in Developmental Education (DevEd) courses has been a growing area of research. Corte and Baptista (2024) investigate this by examining linguistic features and using machine learning to predict placements based on English writing proficiency. This study identifies key linguistic features by analyzing 100 essays, using tools such as COH-METRIX, the Common Text Analysis Platform (CTAP), and manual annotations, revealing that the Naive Bayes algorithm with selected features achieves a classification accuracy of up to 81.8%. In comparison, Kotha et al. (2023) address data imbalance issues in AES and find the random forest algorithm to be the most effective, achieving 97.67% accuracy. Similarly, Suriyasat et al. (2023) developed a Thai essay scoring system where XGBoost and LSTM models yield the best results. Finally, Ramalingam et al. (2018) employ linear regression and other techniques for automated essay assessment, showcasing diverse approaches in the field. AES systems aim to address these challenges by leveraging computational power to provide rapid, consistent scoring across large volumes of essays. Boulanger and Kumar (2018) demonstrate that deep learning methods can significantly improve AES systems' accuracy when applied to the Kaggle Automated Student Assessment Prize dataset. However, they note the limitations due to inadequate training data. Similarly, Li et al. (2022) find that a stacking-based ensemble model enhances AES performance, surpassing neural network baselines by 1.0% to 2.8%. Hussein et al. (2019) propose a trait-specific feedback system, leveraging Long Short-Term Memory (LSTM) networks to improve AES accuracy by 4.6%. In contrast, Sharma and Goyal (2020) confirm that ensemble learning techniques outperform traditional machine learning methods for essay classification. Moreover, Filho et al. (2021) achieved near-perfect accuracy in common classes using feature engineering and deep learning. Ramesh and Sanampudi (2022) identify challenges in assessing essay coherence and relevance, underscoring the need for improved evaluation methods.

Recent advancements in machine learning, particularly in deep learning and ensemble methods, have opened new avenues for improving AES performance. Models such as BERT (Bidirectional Encoder Representations from Transformers) have demonstrated remarkable capabilities in understanding contextual nuances in text (Firoozi et al., 2024), while gradient-boosting algorithms like XGBoost have shown strong predictive power across various domains. Neural networks, renowned for their capacity to learn complex patterns, provide a robust addition to the AES toolkit. However, each of these approaches has its strengths and limitations in the context of essay scoring. This study seeks to push the boundaries of AES by investigating the efficacy of a stacking ensemble method that combines the strengths of XGBoost, neural networks, and ridge regression. This approach is motivated by the hypothesis that different models may excel at capturing distinct aspects of essay quality and that a well-designed ensemble could synthesize these insights for more robust scoring. The study also explores integrating BERT-based models into this ensemble to leverage their advanced language understanding capabilities.

The development of AES systems focuses on various methodologies and evaluation metrics, simulating human evaluation processes. Lim et al. (2021) propose a hybrid framework that integrates both style and content analysis for more accurate predictions. The study recommends Quadratic Weighted Kappa (QWK) as a standard metric for assessing the agreement between human and machine raters. Similarly, Hussein et al. (2019) categorize AES systems into handcrafted and automatically featured systems, noting that while these systems reduce marking workloads and maintain scoring consistency, they face challenges in evaluating creativity and handling deceptive inputs. Andersen et al. (2021) propose alternative evaluation methods and introduce a multitask learning neural network model based on DistilBERT, showing improved performance over traditional models. Meanwhile, Chen et al. (2014) suggest a rank-based approach using pairwise learning to outperform previous AES methods across various datasets. Finally, Xu et al. (2024) identify limitations in existing AES systems, particularly in real classroom settings, highlighting the need for improved scalability, accuracy, and functionality. These studies underscore the potential and challenges in advancing AES technologies for educational use.

A key challenge in AES research is balancing different performance metrics. While Root Mean Squared Error (RMSE) measures overall prediction accuracy, metrics like Cohen's Kappa and Quadratic Weighted Kappa offer insights into the model's ability to agree with human raters across different score ranges. Our study pays particular attention to optimizing performance across these diverse metrics, recognizing that the ideal AES system must be accurate but also consistent and fair in its evaluations. Furthermore, we address the critical aspect of model interpretability. In educational contexts, it is not sufficient for a model to provide a score; it must also offer insights into its decision-making process. This transparency is crucial for maintaining the trust of educators and students and for providing constructive feedback to improve writing skills.

The present study aims to contribute significantly to the field of AES in multiple ways. First, the study provides a comprehensive comparison of state-of-the-art machine learning models, such as BERT, XGBoost, and Neural Networks, to evaluate their strengths and limitations across different evaluation metrics, including Root Mean Squared Error (RMSE), Cohen's Kappa, and Quadratic Weighted Kappa. Second, the study explains how to develop and assess an optimized stacking ensemble method that leverages the strengths of diverse models (XGBoost, Neural Networks, and Ridge Regression), focusing on balancing improvements in RMSE against potential trade-offs in Kappa scores. Third, it explores strategies for optimizing ensemble models in AES through advanced feature engineering, hyperparameter tuning, and integrating transformer-based models like BERT, aiming to enhance predictive accuracy and consistency across various essay types. Finally, the study addresses the challenge of maintaining interpretability in complex ensemble models, ensuring that our AES systems remain transparent and understandable.

The findings of this study have broader implications beyond technical advancements in AES, highlighting its pedagogical and practical applications in educational settings. As this study strives to develop more sophisticated AES systems, the ultimate goal remains clear: to provide fair, accurate, and insightful assessments supporting students' English writing skills.

3. Method

This study optimizes AES by using machine learning models, underscoring the value of AES models in providing objective, consistent scoring in high-stakes educational assessments. Additionally, by addressing the distinct challenges of evaluating complex linguistic features in English essays, the study offers insights relevant to practitioners and researchers seeking to improve accuracy and consistency within English language education. Figure 1 shows the AES optimization process used in this study, with each step subsequently described.

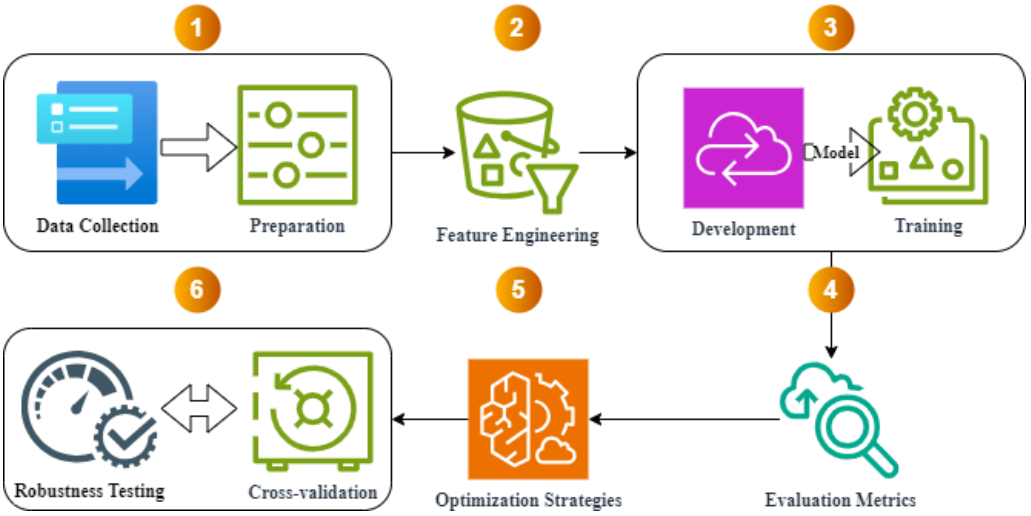

Figure 1. The process of optimizing Automated Essay Scoring (AES) systems by comparing various machine learning models

(Source: Authors, using https://app.diagrams.net)

Step 1: Data Compilation and Preparation

a) Data sources

This study compiled a corpus of English essays written by non-native English speakers, combining two primary datasets: The Automated Student Assessment Prize (ASAP) dataset (Hamner et al., 2012) and the Khon Kaen University Academic English Language Test (KKU-AELT) essays. The ASAP dataset, made available by the Hewlett Foundation through a Kaggle competition, includes approximately 12,976 essays from 7th to 10th-grade students, with lengths ranging from 150 to 550 words per response. Each essay was scored by two raters, with a third rater used if there was significant disagreement, ensuring reliable final scores (Beseiso & Alzahrani, 2020; Yang et al., 2020). However, this dataset has limitations, including varying score ranges and reliance on statistical features for evaluation. The KKU-AELT corpus, obtained with permission from Khon Kaen University's Center for English Language Excellence, consists of 4,817 essays from ten diverse prompts, ranging from 100 to 500 words each. It emphasizes linguistic proficiency among non-native speakers, with each response evaluated by two trained raters, ensuring high interrater reliability (0.80 or higher) (Srisawat & Poonpon, 2023). Integrating the KKU-AELT corpus into the ASAP dataset resulted in a comprehensive combined dataset of 17,793 essays, offering a wide range of essay lengths and types, learners' linguistic backgrounds, and language proficiency (ranging from CEFR A2 to C1). This combined corpus provides a robust foundation for developing AES systems, enhancing objectivity and consistency in evaluation, and allowing researchers to understand the complexities of non-native English writing better.

b) Data score normalization

Merging datasets from different sources introduces potential biases related to writing style, topic familiarity, and scoring criteria. Therefore, the researchers applied normalization techniques to align scoring distributions to address these issues and used stratified sampling for balanced representation. The ASAP dataset utilized a variety of scoring rubrics, with scores ranging from 2 to 12 depending on the prompt, while the KKU-AELT dataset scores ranged from 1 to 5. To align these disparate scales, min-max normalization was applied to rescale all scores to a standard range between 0 and 1, ensuring uniform representation across different essay prompts and datasets. The normalization process followed the formula:

$$S' \frac{S - S_{min}}{S_{max} - S_{min}}$$

Where S represents the original score, S_{min} and S_{max} are the minimum and maximum scores for the respective dataset, and S' is the normalized score. This transformation preserved the relative ranking of essays within each dataset while allowing the model to learn a consistent scoring pattern across datasets.

After normalization, a statistical analysis was conducted to verify that the score distributions across both datasets were sufficiently aligned. Any residual discrepancies were addressed using z-score normalization, ensuring that mean and standard deviation adjustments further

harmonized the scoring scales. This preprocessing step was crucial in mitigating bias introduced by varying scoring rubrics and ensuring model predictions were comparable across datasets.

c) Data cleaning

Another critical step was to standardize textual data by cleaning texts to remove extraneous whitespace and non-ASCII characters, expanding contractions for consistency, and correcting spelling using a dictionary-based approach. Tokenization was also conducted using the WordPiece tokenizer for BERT-based models and standard whitespace tokenization for models like XGBoost, with selective stopwords removal and lemmatization to preserve key linguistic features. Additionally, sentence segmentation and part-of-speech tagging were employed to extract syntactic patterns for feature engineering. All preprocessing steps were implemented using the NLTK and SpaCy libraries, with detailed parameters documented in the supplementary material to ensure reproducibility.

d) Data splitting strategy and generalizability

The dataset was divided into training (70%), validation (15%), and testing (15%) subsets. A 70-15-15 split is a standard practice in machine learning to ensure an optimal balance between training efficiency and model evaluation. The extensive training portion (70%) allows the model to learn robust patterns from a substantial portion of the data, minimizing the risk of underfitting. The 15% validation set was used for hyperparameter tuning and model selection, helping to mitigate overfitting by ensuring that model adjustments were based on an independent subset of data. Finally, the 15% test set provided an unbiased evaluation of the final model's performance on unseen data.

The choice of this split also ensures that the model's generalizability is well assessed. By maintaining a significant portion of data for validation and testing, we can evaluate whether the trained model performs consistently across different essay prompts and scoring variations. Furthermore, the dataset was stratified to preserve the distribution of essay scores across the splits, ensuring each subset contained representative samples of different score ranges. This approach enhances the robustness of the results and provides confidence in the model's ability to generalize beyond the training data. Future studies could experiment with alternative data partitioning strategies, such as k-fold cross-validation, to further assess model stability and mitigate potential biases in score distributions.

Step 2: Feature Engineering

In this study, the feature set comprised three main categories:

a) Linguistic features include word counts, sentence complexity, and vocabulary diversity metrics.

b) Semantic features utilize TF-IDF and word embeddings to capture content relevance.

c) Essay-specific features incorporate domain-specific elements crucial for essay evaluation.

Step 3: Model Development and Training

To facilitate the exploration of the optimizing AES system through various machine learning models, code examples were provided at the beginning of each Python script on our system's web page, hosted on Google Colab (https://colab.research.google.com/drive/1aMhn-WYGjtJngdn46c96No8HdoFUTa_C?usp=sharing).

a) Individual models

The present study employed the BERT-base variant (Bidirectional Encoder Representations from Transformers) developed by Devlin et al. (2019). It built on the transformer architecture introduced by Vaswani et al. (2017), fine-tuning it on the essay dataset. The BERT model leverages a transformer-based approach for natural language processing, using bidirectional training to capture contextual relationships between words within a text. This attention mechanism allows the model to learn complex language patterns effectively. To optimize performance, hyperparameters were adjusted through iterative experimentation, refining the model to better align with the specific characteristics of our dataset.

BERT's architecture allows it to consider the full context of a word by looking at the words that precede and follow it, making it particularly effective for tasks like essay scoring, where understanding context and nuance is crucial.

This initially starts with setting up XGBoost (Extreme Gradient Boosting), an efficient and scalable implementation of gradient boosting machines introduced by Chen and Guestrin (2016) with a broad range of hyperparameters. It is then refined through an extensive grid search process. As an ensemble learning method, XGBoost combines multiple weak learners, usually decision trees, to form a robust predictive model. It builds these trees sequentially, with each new tree specifically designed to correct the errors of the prior ensemble, thereby enhancing overall model performance. XGBoost incorporates several innovations, including a regularized objective, a novel tree-building algorithm, and handling sparse data, making it highly effective for a wide range of machine learning tasks.

A multi-layer neural network architecture was designed and trained using the Adam optimizer with a learning rate of 1e-3. Neural networks, as described by Goodfellow et al. (2016), are computational models inspired by the structure and function of the human brain. These models consist of interconnected nodes, or neurons, organized into an input layer, one or more hidden layers, and an output layer. Each connection between neurons is assigned a weight that is adjusted throughout the learning process, enabling the network to learn complex patterns from data and improve its predictive accuracy over time.

The network learns by adjusting weights and biases to minimize the difference between predicted and actual outputs, typically using

backpropagation and gradient descent algorithms.

Neural networks can be applied to various tasks, including classification, regression, and pattern recognition. Their ability to automatically learn features from data makes them powerful tools in machine learning and artificial intelligence.

b) Stacking ensemble

A stacking ensemble (Wolpert, 1992) combined XGBoost, a Neural Network, and Ridge Regression to enhance prediction performance. In this approach, the meta-learner combines the predictions from the base models and is trained to optimize the final output. Stacking ensembles is an advanced machine-learning technique that leverages the strengths of multiple models to improve overall predictive accuracy. The process involves training several first-level models on the original dataset and using their predictions as inputs to a second-level model or meta-learner. The meta-learner effectively combines the base models' predictions, improving the ensemble's overall performance by minimizing errors and capturing diverse patterns in the data. This approach allows the ensemble to leverage the strengths of diverse models and potentially overcome their weaknesses, often resulting in improved overall performance compared to any single model.

Step 4: Evaluation Metrics

Three primary metrics were used to evaluate writing performance in the English as a Foreign Language (EFL) context. These metrics assess both the accuracy of scoring models and the reliability of human raters.

*a) **Root Mean Square Error (RMSE)***

The RMSE (Chai & Draxler, 2014) was employed to assess the overall prediction accuracy of the model. RMSE is a widely used metric that quantifies the differences between predicted and actual observed values, representing the standard deviation of these residuals or prediction errors. It is beneficial in scenarios where significant errors are undesirable, as it assigns greater weight to more significant discrepancies, thereby providing a more sensitive measure of model performance in accurately predicting outcomes. A lower RMSE indicates that the model's predictions are closer to the actual scores, suggesting better model performance.

*b) **Cohen's Kappa***

Cohen's Kappa (Cohen, 1960) was used to measure inter-rater agreement. This statistical measure assesses the reliability between two raters who classify items into mutually exclusive categories. Unlike a simple percentage agreement, Cohen's Kappa accounts for the likelihood of agreement occurring by chance, providing a more accurate measure of consistency between raters when evaluating categorical data. Cohen's Kappa is beneficial in scenarios where automated systems (e.g., AES) are compared against human raters, as it provides a more robust measure of agreement than simple accuracy metrics.

*c) **Quadratic Weighted Kappa (QWK)***

With its adaptability to different contexts, Quadratic Weighted Kappa (QWK) (Cohen, 1968) was employed to evaluate agreement while considering the ordinal nature of essay scores. QWK, an extension of Cohen's Kappa, is specifically designed for rating scale data, applying weights to disagreements based on their squared distance from perfect agreement. This adaptability makes it particularly useful for measuring agreement between raters or between a rater and a prediction model in contexts like essay scoring, where the magnitude of the difference between scores is important. QWK ranges from -1 to 1, where 1 indicates perfect agreement, 0 indicates agreement equivalent to chance, and negative values indicate agreement less than chance.

These metrics help refine automated scoring systems and ensure consistency among human raters, which is vital for fair and reliable assessments in EFL writing contexts.

Step 5: Optimization Strategies

Feature importance analysis was performed using SHAP (SHapley Additive exPlanations) values for the XGBoost model and attention weights for the BERT model. SHAP values provide a consistent and interpretable measure of each feature's contribution to the model's predictions, highlighting which features are most influential. In contrast, attention weights in BERT indicate the significance of different words or phrases within the input text, revealing how the model focuses on specific elements to understand and predict the target outcomes. This dual approach enables a comprehensive understanding of the importance of features across both traditional machine learning and transformer-based models.

To efficiently explore the parameter space, hyperparameter tuning was conducted using Bayesian optimization (Snoek et al., 2012). Bayesian optimization is a sophisticated method for hyperparameter tuning in machine learning models, which leverages the history of previous evaluations to guide the selection of future hyperparameter values. This approach is more efficient than traditional grid or random search methods, as it can identify optimal configurations more quickly by focusing on the most promising areas of the parameter space.

The core idea is to treat hyperparameter tuning as an optimization problem where we want to find the hyperparameters θ that maximize an objective function f(θ):

$$\theta^* = \text{argmax}_\theta f(\theta)$$

Where f(θ) is typically the validation performance of the model.

Bayesian optimization constructs a probabilistic f(θ) model, usually using Gaussian Processes, and uses an acquisition function to decide which hyperparameter configuration to evaluate next. A common acquisition function is Expected Improvement (EI):

$$EI(\theta) = E[\max(f(\theta) - f(\theta+), 0)]$$

Where θ+ is the best observed hyperparameter configuration so far.

This approach balances exploration (trying new areas of the hyperparameter space) and exploitation (focusing on areas known to perform well), potentially leading to more efficient hyperparameter tuning.

The ensemble weights were optimized to balance the contributions of each model to the final prediction. In machine learning, ensemble weights determine the relative importance of each model within an ensemble, influencing how their predictions are combined to produce a final output (Zhou, 2012). By assigning appropriate weights, the ensemble can capitalize on the strengths of each model while mitigating their weaknesses, thereby significantly enhancing the overall predictive performance. This weighted approach ensures that the final prediction benefits from the unique capabilities of each model in the ensemble.

Step 6: Cross-validation and Robustness Testing

A 5-fold cross-validation strategy was implemented to ensure model stability and generalizability across different data subsets. This technique involves partitioning the dataset into five equally sized subsets or folds. The model is then trained on four folds and validated on the remaining fold, with this process repeated five times so that each fold serves as the validation set once. This approach provides a more reliable estimate of the model's performance on unseen data compared to a single train-test split, as it utilizes all observations for both training and validation and mitigates the impact of data partitioning on the results (Kohavi, 1995).

In complementing the quantitative analysis, a qualitative analysis was conducted on sample essays scored by various models to assess their ability to capture the nuances of essay quality. This involved examining a selection of high-scoring and low-scoring essays to determine the models' effectiveness in recognizing linguistic complexity, coherence, argument structure, and vocabulary use.

4. Results

The results highlight the varying performance of different models for automated essay scoring. The BERT-based transformer model showed consistent improvement over three training epochs, with the training loss decreasing from 314.60 in Epoch 1 to 172.05 in Epoch 3 and the validation loss reducing from 457.10 to 346.49 over the same period. The final performance metrics for the BERT model indicated a Root Mean Squared Error (RMSE) of 18.61, Cohen's Kappa of 0.0120, and a Quadratic Weighted Kappa (QWK) of 0.256. In contrast, the initial XGBoost model demonstrated an RMSE of 20.48, a Cohen's Kappa of 0.0038, and a QWK of 0.104. After undergoing hyperparameter optimization, the XGBoost model achieved an improved RMSE of 17.99 and a QWK of 0.204, utilizing parameters such as a colsample_bytree of 0.8, a learning rate of 0.01, a max depth of 4, 300 estimators, and a subsample of 1.0. The stacking ensemble, which combined XGBoost, a Neural Network, and Ridge Regression, resulted in a substantially lower RMSE of 1.84, though with a slight reduction in agreement metrics, yielding a Cohen's Kappa of -0.0075 and a QWK of 0.175. A comparative analysis showed that while the stacking ensemble achieved a significant reduction in RMSE, lowering it from 17.99 (the best individual model) to 1.84, this improvement came with a trade-off in the QWK, which decreased from 0.256 (achieved by the BERT model) to 0.175.

Table 1. A comparison of the results of different models used in the study

Model	RMSE	Cohen's Kappa	Quadratic Weighted Kappa
BERT-based Transformer	18.61	0.0120	0.256
XGBoost (Initial)	20.48	0.0038	0.104
XGBoost (Optimized)	17.99	N/A	0.204
Stacking Ensemble	1.84	-0.0075	0.175

*Note: The Cohen's Kappa for the optimized XGBoost model was not provided in the original data.

Table 1 shows that the Stacking Ensemble achieved the best RMSE of 1.84, demonstrating a substantial improvement in predictive accuracy compared to other models. However, this approach slightly decreased the QWK, highlighting a trade-off between accuracy and agreement metrics. In contrast, the BERT-based Transformer model attained the highest QWK score of 0.256, suggesting its superior performance in maintaining consistency across different scoring ranges. Meanwhile, the XGBoost model's performance improved markedly following optimization, with notable gains in both RMSE and QWK, indicating its enhanced ability to balance predictive accuracy and reliability.

These results highlight the complex nature of essay scoring, where improvements in one metric may not necessarily translate to improvements across all evaluation criteria. The dramatic reduction in RMSE suggests that the ensemble model provides more precise score predictions on average. However, the decrease in Kappa scores indicates potential issues with consistency across different score ranges or essay types.

This trade-off underscores the crucial role of a nuanced approach in model selection and optimization for automated essay scoring. It is not just about overall accuracy but also about consistency and alignment with educational assessment principles. Your expertise and careful consideration in this process are paramount.

5. Discussion

The study's findings on optimizing Automated Essay Scoring (AES) systems through the ensemble methods offer several insights and highlight critical areas for future exploration in this field. Initially, our investigation focused on evaluating individual models, including the BERT-based Transformer, XGBoost, and Neural Network. The BERT-based model emerged with the highest Quadratic Weighted Kappa (QWK) score of 0.256, demonstrating its potential to capture nuanced linguistic features and contextual information essential for effective essay scoring. This finding aligns with Lim et al. (2021), who emphasized the capability of machine learning models to mirror human scoring more closely by understanding complex language patterns. The initial performance of the XGBoost model, with a lower QWK of 0.104 and an RMSE of 20.48, nevertheless provided valuable insights into the importance of feature engineering and hyperparameter tuning. After optimization, the XGBoost model exhibited substantial improvement (RMSE: 17.99; QWK: 0.204), illustrating the model's sensitivity to its configuration and underscoring the potential for significant enhancement through careful tuning, as also observed by Hussein et al. (2019), who highlighted the critical role of feature engineering in AES development.

This study's most notable result was the stacking ensemble's performance, which combined XGBoost, Neural Network, and Ridge Regression models. This ensemble approach led to a significant reduction in RMSE, from 17.99 in the best individual model to 1.84 in the ensemble, indicating that the ensemble was highly effective in minimizing the average magnitude of scoring errors and potentially offering more precise score predictions overall. However, this improvement in RMSE came at a cost. The ensemble model displayed a slight decrease in QWK, with a score of 0.175 compared to the BERT model's 0.256 and even the optimized XGBoost model's 0.204. This reduction, coupled with a negative Cohen's Kappa of -0.0075, raises important questions about the nature of the ensemble's predictions and their consistency with human raters, similar to the challenges noted by Andersen et al. (2021), who discussed difficulties in achieving alignment between AES models and human examiners.

While the stacking ensemble achieved a significantly lower RMSE, the persistent negative Cohen's Kappa (-0.0075) suggests a fundamental misalignment between the predicted scores and human raters. This suggests that the model's predictions are worse than random chance and raises concerns about systematic biases in the scoring process. Unlike RMSE, which measures numerical accuracy, Cohen's Kappa assesses inter-rater agreement by comparing the observed agreement to what might occur by chance; thus, the negative value highlights alignment issues in the ensemble. The stacking ensemble, which optimizes RMSE by aggregating multiple models such as XGBoost, Neural Network, and Ridge Regression, may have inadvertently emphasized inconsistencies, overfitting to statistical patterns misaligned with human grading. This could result in systematic biases where certain essay features are disproportionately weighted, skewing predictions away from human expectations. Furthermore, imbalances in data distribution could lead to inaccurate decision boundaries, causing erratic predictions across score levels. Differences in rubric interpretation between datasets, such as ASAP and KKU-AELT, might exacerbate inconsistencies in scoring criteria.

The ensemble model demonstrated a trade-off between RMSE and QWK scores, revealing a critical challenge in AES. While achieving a significantly reduced RMSE (1.84 from 17.99), indicating high numerical accuracy, the model exhibited lower Kappa scores (QWK 0.175; Cohen's Kappa decreased by 0.0075), suggesting potential consistency or bias issues. This divergence, consistent with Xu et al. (2024), highlights limitations in meeting diverse classroom needs. Factors such as Kappa score sensitivity, QWK error penalties, and model calibration may contribute. The trade-off underscores the difficulty of balancing numerical accuracy (RMSE) with ordinal coherence and inter-rater reliability (Kappa). High-stakes assessments prioritize QWK for ranking consistency, while formative assessments may favor RMSE for precise feedback.

While Large Language Models (LLMs), e.g., GPT-4 and fine-tuned GPT-3.5, demonstrate strong performance in AES through their strong natural language understanding and scoring automation capabilities (e.g., Pack et al., 2024), this study utilizes a stacking ensemble prioritizing interpretability. Unlike LLMs, the ensemble method offers transparency through SHAP-based feature analysis and model explainability, enabling educators and stakeholders to trace score derivations. Additionally, it provides a more accessible alternative for large-scale assessments. Although LLMs excel in accuracy and generalizability, they pose challenges in interpretability, computational expense, and potential bias.

In complementing the quantitative findings, the qualitative result reveals distinct strengths and weaknesses in each model. The BERT-based model excelled at recognizing coherence, sophisticated vocabulary, varied sentence structures, and logical progression but sometimes overemphasized lexical sophistication. The XGBoost model effectively differentiated essays based on basic readability metrics, such as sentence length and word frequency, yet it struggled with nuanced rhetorical structures and reasoning. The stacking ensemble model, which merged multiple methodologies, generally achieved balanced scoring but occasionally prioritized specific linguistic features, such as sentence complexity, over overall coherence. While machine learning models efficiently automate essay scoring, they still struggle to fully replicate human judgment (Toranj & Ansari, 2012; Zribi & Smaoui, 2021).

From an English language teaching perspective, this study suggests that an ensemble AES system, integrating models, i.e., BERT, XGBoost, and neural networks, effectively supports English language learners. Improved accuracy (low RMSE) ensures fair and precise evaluation, mitigating linguistic or cultural biases. High-quality, interpretable feedback via SHAP values and attention mechanisms fosters independent

learning and skill development in areas like coherence and argumentation. Scalable assessments allow teachers to provide individualized support, which is crucial for large ESL/EAP programs. Harshalatha and Sreenivasulu (2024) highlight the importance of practice and motivation in ESL writing, reinforcing the study's contribution to improving writing proficiency and supporting non-native speakers in achieving academic success.

Additionally, this study reveals the effectiveness of ensemble-based AES models, particularly those utilizing BERT, in enhancing high-stakes English language proficiency tests. Advanced transformer models effectively capture linguistic nuances, enabling accurate assessment of advanced skills such as nuanced vocabulary, grammar, and argument coherence (Firoozi et al., 2024; Settles et al., 2020). Integrating RMSE with QWK aligns automated scores with human judgment, ensuring scores reflect true language proficiency. These models comprehensively evaluate linguistic and argumentative skills by incorporating essay-specific features like sentence complexity and vocabulary diversity. Interpretability within AES models allows for transparent feedback, benefiting test-takers and instructors (Wei et al., 2023).

6. Implications of the Study

The present study provides significant implications across methodological, assessment, and pedagogical dimensions. For methodological implications, the ensemble model highlights the importance of choosing an appropriate evaluation metric, as it can influence conclusions about model performance. While RMSE provides insights into prediction accuracy, metrics like QWK are essential for ensuring consistency with human raters, as emphasized by Lim et al. (2021). The ensemble's ability to reduce RMSE suggests that different models can capture unique aspects of essay quality, indicating that future AES systems should incorporate a range of model architectures to assess essays comprehensively. However, optimizing both accuracy (RMSE) and consistency (Kappa scores) simultaneously remains a challenge, reflecting the complexities noted by Hussein et al. (2019) in balancing these objectives.

In the assessment dimension, this study demonstrates the efficacy of ensemble-based AES models, especially those utilizing BERT, in refining high-stakes English language proficiency tests (Pack et al., 2024). These systems can be used to assess English language learners' writing skills more accurately, which can help them improve their proficiency and achieve their academic goals. The interpretability of these systems also allows test-takers to see how their scores were calculated and identify areas for improvement. This feedback can also be helpful for teachers in providing targeted instruction to improve their students' proficiency (Guo, 2023). These AES systems can effectively assess nuanced vocabulary, grammar, and argument coherence. This means that they can be used to assess a variety of writing tasks, such as essays, letters, and reports.

The pedagogical implications of this study for English language teaching and assessment are multifaceted. First, AES systems should be viewed as complementary tools to human scoring, particularly useful for providing rapid feedback in formative assessments or as a preliminary step in large-scale assessments (Wei et al., 2023; Zribi & Smaoui, 2021). However, continuous evaluation and refinement are necessary to ensure their effectiveness (Andersen et al., 2021; Lim et al., 2021; Chansanam et al., 2021). AES systems need to be tailored to specific educational contexts and essay types to maximize their utility. EFL teachers require professional development to understand the capabilities and limitations of these systems, ensuring they interpret outputs critically and effectively and optimize the use of these technologies. For instance, SHAP-based feedback can enhance learners' writing skills by providing detailed insights into their performance. It can also be tailored to support students from diverse proficiency levels, promoting equitable and personalized learning. Moreover, linking AES systems to authentic academic writing tasks, such as argument and summary writing, can bridge the gap between research and classroom practice.

7. Limitations and Suggestions for Future Research

This study highlights some limitations that offer promising avenues for future research. First, the dataset's specific characteristics may influence model performance, so future studies should investigate the generalizability across essay types, levels, and scoring rubrics (Xu et al., 2024). Second, incorporating more advanced NLP techniques or domain-specific features could enhance model performance (Chen et al., 2014). Furthermore, improving the interpretability of complex AES systems, potentially through hybrid systems combining AI predictions with human expertise, is crucial for transparent assessments (Xu et al., 2024). To address the misalignment between the model's predictions and human raters, future research should explore bias mitigation via reweighting to improve agreement with human raters, ordinal regression to maintain better score rankings, and performing error analyses to identify and correct specific misclassification patterns. Post-hoc calibration, e.g., Platt scaling or isotonic regression, can better align predictions with human scoring distributions. The present study also highlights the RMSE and Kappa trade-off. Future research should explore hybrid strategies, e.g., ordinal regression and calibration techniques, to better align numerical accuracy with categorical consistency. Addressing these trade-offs can ultimately enhance fairness and reliability in AES systems.

To fully leverage AES systems in EFL settings, further research is necessary to understand how teachers integrate these tools into their teaching practices and how automated feedback influences student writing development. Conducting surveys or interviews with EFL teachers can provide insights into their experiences with AES systems, highlighting common challenges, benefits, and areas for improvement in teacher training and support. Additionally, longitudinal studies examining the long-term effects of automated feedback on student writing skills would offer valuable evidence for educators and policymakers. By addressing these research gaps, we can ensure that AES systems are used effectively to enhance teaching and learning outcomes in EFL contexts.

8. Conclusion

This study has explored optimizing Automated Essay Scoring (AES) systems through a comparative analysis of machine learning approaches, focusing on ensemble methods. The research has yielded several significant findings that contribute to the advancement of AES technology and its application in educational contexts. First, the comparative analysis of individual models—BERT, XGBoost, and Neural Networks—revealed each approach's unique strengths and limitations in English essay scoring. The BERT-based model demonstrated superior performance in capturing nuanced linguistic features and context, while XGBoost efficiently handled diverse feature sets. Neural Networks showed promise in learning complex patterns within the essay data. The core study—developing and evaluating a stacking ensemble method—marked a significant advancement in AES performance. By integrating XGBoost, Neural Networks, and Ridge Regression, this study substantially reduced Root Mean Squared Error (RMSE), decreasing it from 17.99 in the best individual model to 1.84 in the ensemble. This notable reduction in RMSE signifies a considerable improvement in the accuracy of essay score predictions.

The findings also highlighted the complexities of optimizing AES systems. The slight decrease in Quadratic Weighted Kappa from 0.204 to 0.175 in the ensemble model underscores the challenges in balancing different performance metrics. This trade-off between RMSE and Kappa scores emphasizes the need for a nuanced approach to model evaluation in educational applications, where consistency across score ranges is as crucial as overall accuracy. Notably, this research made significant strides in addressing model interpretability by implementing feature importance analysis and utilizing SHAP (SHapley Additive exPlanations) value. This advancement is crucial for practical applications in educational settings, as understanding the rationale behind scores is essential for both teachers and students. Integrating BERT into the ensemble model represents a promising direction for future research, although challenges remain in fully leveraging transformer-based models within an ensemble framework. The initial results of the present study suggest that this integration could further enhance accuracy and linguistic understanding. Furthermore, this study emphasizes the importance of comprehensive feature engineering in AES. Developing domain-specific and advanced linguistic and semantic features was pivotal in improving model performance, highlighting the need for a deep understanding of both the technical aspects of machine learning and the pedagogical principles of essay assessment.

In conclusion, the stacking ensemble method in the present study has demonstrated substantial potential in enhancing AES performance, particularly in reducing RMSE. However, it also reveals the complexities of essay scoring, with slight trade-offs in other metrics indicating opportunities for further refinement. Future research should prioritize refining ensemble techniques to better balance performance metrics, enhance interpretability, and explore innovative ways to integrate advanced language models such as BERT into ensemble frameworks. As AES systems continue to evolve, it is essential to balance advancing technical capabilities and ensuring these systems fulfill their educational purpose. The ultimate goal is to develop AES tools that provide accurate scores and offer meaningful insights to support students' writing development. This research contributes to this ongoing endeavor, laying the groundwork for more sophisticated, equitable, and educationally valuable automated essay scoring systems.

Acknowledgments

The authors would like to thank the Center for English Language Excellence, Faculty of Humanities and Social Sciences, Khon Kaen University, Thailand, for providing the essay dataset used in this study.

Authors' contributions

Dr. Kornwipa and Dr. Wirapong designed the study. Dr. Kornwipa handled data collection, while Dr. Wirapong conducted the data analysis. Dr. Wirapong drafted the manuscript, and Dr. Kornwipa revised it. All authors reviewed and approved the final manuscript and contributed equally to the study.

Funding

This work was supported by the Fundamental Fund of Khon Kaen University. The research titled "A Comparative Study of English as a Foreign Language Learners' Academic Writing Ability Evaluated by Learning Algorithm and Human Rater Judgment" also received funding from the National Science, Research, and Innovation Fund (NSRF).

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Obtained.

Ethics approval

The authors declare that an ethical review and approval were waived for this study, as it used pre-existing data that is openly accessible and did not require approval from an ethics committee.

The Publication Ethics Committee of the Sciedu Press.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

References

Andersen, Ø. E., Yuan, Z., Watson, R., & Cheung, K. Y. F. (2021). Benefits of alternative evaluation methods for automated essay scoring. *International Educational Data Mining Society.* Retrieved from https://educationaldatamining.org/EDM2021/virtual/static/pdf/EDM21_paper_179.pdf

Bachman, L., & Palmer, A. S. (1996). *Language testing in practice.* Oxford University Press.

Bejar, I. I. (2011). A validity-based approach to quality control and assurance of automated scoring. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 319-341. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/0969594X.2011.555329

Beseiso, M., & Alzahrani, S. (2020). An empirical analysis of BERT embedding for automated essay scoring. *International Journal of Advanced Computer Science and Applications*, *11*(10), 204-210. https://doi.org/10.14569/IJACSA.2020.0111027

Boulanger, D., & Kumar, V. (2018). Deep learning in automated essay scoring [Review of *Deep learning in automated essay scoring*]. In *Intelligent Tutoring Systems: 14th International Conference*, Vol. 14 (pp. 294-299). Springer International Publishing. https://doi.org/10.1007/978-3-319-91464-0_30

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014, 2014.

Chansanam, W., Poonpon, K., Manaku, T., & Detthamrong, U. (2021). Success and challenges in MOOCs: A literature systematic review technique. *TEM Journal*, 10(4), 1728–1732. https://doi.org/10.18421/TEM104-32

Chen, H., Xu, J., & He, B. (2014). Automated essay scoring by capturing relative writing quality. *The Computer Journal*, *57*(9), 1318-1330. https://doi.org/10.1093/comjnl/bxt117

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213-220. https://doi.org/10.1037/h0026256

Corte, M. D., & Baptista, J. (2024). Leveraging NLP and machine learning for English (L1) writing assessment in developmental education. *International Conference on Computer Supported Education*. https://doi.org/10.5220/0012740500003693

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & Jamse, M. (2005). Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL®. *ETS Research Report Series*, *2005*(1), i-77. https://doi.org/10.1002/j.2333-8504.2005.tb01990.x

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers) (pp. 4171-4186).

Filho, A. H., Concatto, F., Antonio do Prado, H., & Ferneda, E. (2021). Comparing feature engineering and deep learning methods for automated essay scoring of Brazilian National High School Examination. In *Proceedings of the 23rd International Conference on Enterprise Information Systems*, Vol. 1 (pp. 575-583). ICEIS. https://doi.org/10.5220/0010377505750583

Firoozi, T., Mohammadi, H., & Gierl, M. J. (2024). Using automated procedures to score educational essays written in three languages. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12406

Geçkin, V., Kızıltaş, E., & Çınar, Ç. (2023). Assessing second-language academic writing: AI vs. human raters. *Journal of Educational Technology & Online Learning*, *6*(4), 1096-1108. https://doi.org/10.31681/jetol.1336599

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning. MIT Press*. Retrieved from http://www.deeplearningbook.org

Green, A. (2022). *L2 Writing Assessment: An evolutionary perspective*. Palgrave Macmillan Cham. https://doi.org/10.1007/978-3-031-15011-1

Guo, T. (2023). The impact of written corrective feedback on enhancing second language writing skills: A comprehensive review. *International Journal of Education and Humanities (IJEH)*, *3*(3), 280-287. https://doi.org/10.58557/(ijeh).v3i3.132

Hamner, B., Morgan, J., Lynnvandev, Shermis, M., & Ark, T. V. (2012). *The Hewlett Foundation: Automated essay scoring*. Retrieved from https://www.kaggle.com/c/asap-aes/overview/ evaluation

Hamp-Lyons, L., & Kroll, B. (1997). *Assessing Second Language Writing*. Cambridge University Press.

Harshalatha, S., & Sreenivasulu, Y. (2024). Exploring academic writing needs and challenges experienced by ESL learners: A literature review. *World Journal of English Language*, *14*(3), 406-406. https://doi.org/10.5430/wjel.v14n3p406

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, *5*, e208. https://doi.org/10.7717/peerj-cs.208

Hyland, K. (2003). *Second Language Writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511667251

Klebanov, B. B., & Madnani, N. (2022). *Automated essay scoring*. Springer Nature. https://doi.org/10.1007/978-3-031-02182-4

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings 14th International Joint Conference Aritificial Intelligence (IJCAI)*, Vol. 2 (pp. 1137-1145). Retrieved from https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf

Kotha, U. M., Gaddam, H., Siddenki, D. R., & Saleti, S. (2023). A comparison of various machine learning algorithms and execution of flask deployment on essay grading. *International Journal of Electrical and Computer Engineering (IJECE)*, *13*(3), 2990-2998. http://doi.org/10.11591/ijece.v13i3.pp2990-2998

Lahitani, A. R., Permanasari, A. E., & Setiawan, N. A. (2016). Cosine similarity to determine similarity measure: Study case in online essay assessment. *The 4th International Conference on Cyber and IT Service Management*, 1-6. https://doi.org/10.1109/CITSM.2016.7577578

Li, C., Lin, L., Mao, W., Xiong, L., & Lin, Y. (2022). An automated essay scoring model based on stacking method. In *2022 IEEE 2nd International Conference on Software Engineering and Artificial Intelligence (SEAI)* (pp. 248-252). https://doi.org/10.1109/SEAI55746.2022.9832246

Lim, C. T., Bong, C. H., Wong, W. S., & Lee, N. K. (2021). A comprehensive review of automated essay scoring (AES) research and development. *Pertanika Journal of Science & Technology*, *29*(3), 1875-1899. https://doi.org/10.47836/pjst.29.3.27

Manipatruni, V. R., Kumar, N. S., Karim, M. R., & Banu, S. (2024). Improving writing skills through essay writing via 'Write & Improve' for error analysis and 'Padlet' for collaborative writing & peer review. *World Journal of English Language*, *14*(4), 204-214. https://doi.org/10.5430/wjel.v14n4p204

Oshima, A., & Hogue, A. (2007). *Introduction to academic writing*. Pearson Education.

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, *6*. https://doi.org/10.1016/j.caeai.2024.100234

Ramalingam, V., Pandian, A. A., Chetry, P., & Nigam, H. (2018). Automated essay grading using machine learning algorithm. *Journal of Physics: Conference Series*, *1000.* https://doi.org/10.1088/1742-6596/1000/1/012030

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495-2527. https://doi.org/10.1007/s10462-021-10068-2

Settles, B., Hagiwara, M., & LaFlair, G. T. (2020). Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, *8*, 247-263. https://doi.org/10.1162/tacl_a_00310

Sharma, S., & Goyal, A. (2020). Automated essay grading: An empirical analysis of ensemble learning techniques. In Singh, V., Asari, V. K., Kumar, S., Patel, R.B. (Eds.), *Computational Methods and Data Engineering. Advances in Intelligent Systems and Computing*, Vol. 1257. Springer, Singapore. https://doi.org/10.1007/978-981-15-7907-3_26

Shermis, M., & Wilson, J. (2024). *The Routledge international handbook of automated essay evaluation.* Routledge. https://doi.org/10.4324/9781003397618

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems* (pp. 2951-2959). https://doi.org/10.48550/arXiv.1206.2944

Srisawat, C., & Poonpon, K. (2023). Revision of an academic English writing rubric for a graduate school admission test. *PASAA*, *65*, 234-262. https://doi.org/10.58837/CHULA.PASAA.65.1.9

Suriyasat, S., Chanyachatchawan, S., & Tuaycharoen, N. (2023). A Comparison of machine learning and neural network algorithms for an

automated Thai essay scoring. *20th International Joint Conference on Computer Science and Software Engineering (JCSSE)* (pp. 55-60). https://doi.org/10.1109/JCSSE58229.2023.10201964

Swales, J. M., & Feak, C. B. (2012). *Academic writing for graduate students: Essential tasks and skills*. University of Michigan Press. https://doi.org/10.3998/mpub.2173936

Toranj, S., & Ansari, D. N. (2012). Automated versus human essay scoring: A comparative study. *Theory and Practice in Language Studies*, *2*(4), 719-725. https://doi.org/10.4304/tpls.2.4.719-725

Tracy-Ventura, N., & Paquot, M. (Eds.) (2020). *The Routledge handbook of second language acquisition and corpora*. Routledge. https://doi.org/10.4324/9781351137904

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008). Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wei, P., Wang, X., & Dong, H. (2023). The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: A randomized controlled trial. *Frontiers in Psychology*. *14*, 1249991. https://doi.org/10.3389/fpsyg.2023.1249991

Weigle, S. (2002). *Assessing writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1

Xiao, C., Ma, W., Xu, S., Zhang, K., Wang, Y., & Fu, Q. (2024). From automation to augmentation: Large language models elevating essay scoring landscape. *ArXiv*, abs/2401.06431.

Xu, W., Mahmud, R. B., & Lam, H. W. (2024). A systematic literature review: Are automated essay scoring systems competent in real-life education scenarios? *IEEE Access*, *12*, 77639-77657. https://doi.org/10.1109/ACCESS.2024.3399163

Yang, R., Cao, J., Wen, Z., Wu, Y., & He, X. (2020). Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1560-1569). https://doi.org/10.18653/v1/2020.findings-emnlp.141

Zhang, C. (2025). A brief review of recent research on second-language writing. *Lecture Notes in Education Psychology and Public Media*, *78*(1), 79-83. https://doi.org/10.54254/2753-7048/2025.19169

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R&D Connections*, *21*(2), 1-11.

Zhou, Z. H. (2012). *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC. https://doi.org/10.1201/b12207

Zribi, R., & Smaoui, C. (2021). Automated versus human essay scoring: A comparative study. *International Journal of Information Technology and Language Studies*, *5*(1), 62-71. Retrieved from https://journals.sfu.ca/ijitls/index.php/ijitls/article/view/199