Deep Learning and Rule-Based Hybrid Model for Enhanced English Composition Scoring Using Attention Mechanisms and Graph Convolutional Networks

Ruimin Li
Zhoukou Vocational and Technical College, Zhoukou 466000, China
E-mail: laogui9029@126.com

Keywords: English essay grading, deep learning, artificial rules, graph convolutional network, wide&deep architecture

Technical paper

Received: July 8, 2025

Through the profound exploration conducted on AI technology in the field of education, early automatic scoring systems for English compositions have problems such as high misjudgment rate and low efficiency. To improve the efficiency, accuracy, and stability of the English composition grading model, a deep learning and manual rule-based English composition grading model was designed. The research extracted sequence features by introducing attention mechanisms, enhancing contextual correlation analysis, and aggregating global features through graph convolutional networks to extract high-order semantic relationships. Finally, a visual manual scoring rule was designed, which integrated deep semantic features and manual rule features through the Wide&Deep architecture to jointly optimize the scoring results. The experiment outcomes indicated that the accuracy recall curve area of the research method was 92.3%. In practical application testing, the highest group stability index of the research method was 0.07 in June. When faced with 600 concurrent requests, the average response time of the research method reached a stable value of 3.4 seconds. The outcomes above demonstrated that the English essay scoring model, which combines deep learning with manual rules as proposed by the research, exhibited excellent accuracy, speed, and stability. It effectively addressed the issues of a high misjudgment rate and low efficiency found in traditional scoring systems, thereby enhancing the model's reliability.

Povzetek: Razvit je hibridni model za ocenjevanje angleških esejev, ki združuje globoko učenje z ročnimi pravili. Z Word2Vec, mehanizmom pozornosti in GCN zajame lokalne ter globalne semantike, Wide&Deep pa združi pravila in značilke.

1 Introduction

English writing ability is one of the core indicators of language learning, and traditional manual scoring methods face bottlenecks such as low efficiency and strong subjectivity [1, 2]. Early automatic scoring systems mainly relied on rule-based methods to detect surface errors through pre-defined grammar and spelling rules, but it was difficult to evaluate the quality of content and logic, resulting in a high rate of misjudgment [3]. With the advancement of technology, machine learning (ML) algorithms have been introduced to comprehensively consider vocabulary, syntax, and other elements through feature engineering. However, a substantial quantity of annotated data support is still needed, and the generalization ability is insufficient [4]. The existing scoring system cannot meet the automatic scoring requirements for English compositions, and there is an urgent need for a stable, efficient, and accurate scoring model. Deep learning (DL) models can improve semantic understanding through end-to-end learning, but they lack transparency and are difficult to capture grammatical details. Artificial rules have a high degree of interpretability, but cannot adapt to open content evaluation. The two methods complement each other in advantages [5]. In light of the preceding circumstances, to ensure the stability, accuracy, and efficiency of the scoring model, an innovative English composition scoring model based on DL and artificial rules has been designed. The research uses Word2Vec model to convert essay text into a matrix of word vectors, capturing semantic information of vocabulary. It introduces attention mechanism and graph convolutional network to extract local sequence features and semantic graph features, and concatenating the two features to generate deep semantic features, constructing graph adjacency matrix to dynamically capture the relationships between sentences. Then, artificial rule features are generated through feature concatenation, and the Wide&Deep architecture is used to fuse deep semantic features with artificial rule features. Finally, combining multi-dimensional manual rule evaluation, the research achieves dynamic comprehensive scoring of the entire English composition. It is anticipated that research methodologies will offer a theoretical foundation for grading essays in different languages.

2 Related works

English composition grading is an important part of the educational evaluation system, playing a crucial role in achieving teaching objectives and optimizing teaching strategies. Ramesh et al. proposed AI and ML techniques for evaluating automatic paper grading in response to issues such as time-consuming manual assessments and lack of reliability in the education system. During the research process, the limitations and research trends of the current study were analyzed. The outcomes revealed that the research method had a good effect [6]. Fokides et al. compared the accuracy and qualitative aspects of the corrections and feedback generated by ChatGPT with educators regarding the effectiveness of ChatGPT on elementary school students' essays written in English. The outcomes revealed that ChatGPT surpassed educators in regard to both the volume and the caliber of output [7]. Shahzad et al. proposed using random forests as classifiers for off topic paper detection to address the prediction problem of whether an article deviates from the topic. The outcomes revealed that the research method had high accuracy [8]. Erturk et al. pointed out the low reliability and effectiveness of essay style evaluation tools, and believed that the system's decrease in paper scores was related to boredom in the labeling. The outcomes revealed that higher levels of boredom were correlated with lower scores [9]. Sharma et al. proposed a system that combines handwriting recognition models and automatic paper grading to address the time-consuming issue of grading handwritten papers in educational environments. During the research process, the performance of downstream tasks in paper scoring was analyzed based on Transformer context embedding. The outcomes revealed that the research method had good performance [10].

Many scholars both within the country and abroad have carried out profound investigations and application of Word2Vec and artificial rules. Mohammed et al. conducted an exhaustive examination of diverse approaches within the realm of ensemble learning to address the issue of time-consuming hyperparameter tuning in DL. Various features or factors that affect the success of integration methods were explained during the research process. The outcomes revealed that the research method could provide accurate theoretical support [11]. Tropsha et al. proposed a "deep quantitative structure-activity relationship" model for virtual screening of molecular databases. The outcomes revealed that the research method had a good effect [12]. Whang et al. proposed a fairness measure and unfairness mitigation technique to address the issues of bias and unfairness in traditional data management. The outcomes revealed that the research method had good data management performance [13]. Pereira et al. proposed an ML system for multi animal pose tracking to address the challenge of using DL and computer vision techniques to study the social behavior of multiple animals in natural environments. The outcomes revealed that the research method had good efficiency and accuracy [14]. Olan et al. designed an explanatory algorithm to address the impact of AI on the decision-making process in the supply chain field. The composition of interpretable AI and decision support systems was determined during the research process. The outcomes revealed that the research method could effectively enhance decision-making ability in the context of supply chain [15].

In summary, existing research has played a good role in the technological advancement of English composition grading models, but it still has limitations such as low grading efficiency and significant subjective differences. The automatic scoring model based on DL can extract multi-level information such as linguistic features and semantic information, which can simulate the process of manual scoring to a certain extent, while manual rules can handle complex grammar rules and subtle semantic differences. Therefore, based on this, a DL and artificial rule-based English composition grading model was designed. The goal is to align with the design standards for automated English composition grading and to significantly boost both the accuracy and efficiency of the grading workflow.

3 Design of english composition scoring model

3.1 Intelligent english composition scoring model based on deep semantic text features

As an important part of the education evaluation system, English composition grading has undergone an evolution from traditional manual grading to automated grading. However, existing automated grading systems are mostly based on shallow text features, resulting in significant errors in their grading results [16, 17]. DL models can effectively improve the accuracy and reliability of English composition grading from three aspects: feature extraction, semantic understanding, and grading prediction [18]. The study converts the original English composition text into a numerical word embedding matrix, and the text to word embedding conversion formula is shown in Equation (1).

$$\mathbf{E} = \text{Embedding}(\mathbf{X}) \quad (1)$$

In Equation (1), $\mathbf{X}$ represents the input English composition text sequence, represents the word embedding matrix, and $\text{Embedding}()$ represents a DL embedding function. Next, the research will investigate the use of the Woed2Vec learning model to map each word to a high-dimensional space, capturing the semantic and positional information of the word. The Word2Vec learning model has two training models: continuous bag of words and skip word. The frameworks of the two models are presented in Figure 1.

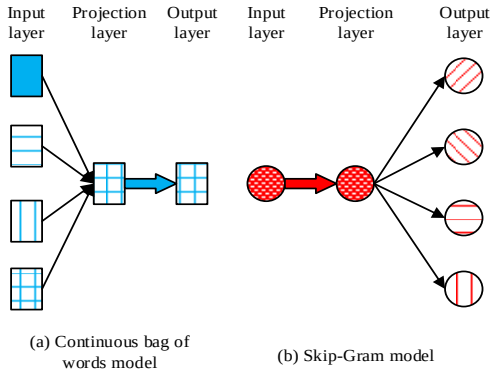


Figure 1: Framework diagram of continuous bag of words model and skip-gram model

As shown in Figure 1, the training process of both the continuous bag of words training model and the skip word training model goes through the input layer, the mapping layer, and finally outputs the results from the output layer. However, the continuous bag of words model aggregates and maps multiple features, and then outputs the results, while the skip word model maps the features and performs classification output. The study combines continuous bag of words training model and skip word training model to train and detect the sequence features and semantic map features of English compositions, and then scores the English compositions based on the detection results. In the process of extracting sequence features from English compositions, in order to break through the sequence limitations of DL models, self-attention mechanism is introduced, and its function expression is shown in Equation (2).

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{2}$$

In Equation (2), $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the query matrix, key matrix, and value matrix, respectively. d_k is the dimension of the key or query vector and $\text{softmax}()$ represents the normalization function. To enhance the model's ability to express complex sequence patterns, a multi-head attention mechanism is introduced, and its calculation formula is shown in Equation (3).

$$\begin{cases} \text{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)\mathbf{W}^O \\ \text{wherehead}_i = \text{Attention}(\mathbf{X}\mathbf{W}_i^Q,\mathbf{X}\mathbf{W}_i^K,\mathbf{X}\mathbf{W}_i^V) \end{cases} \tag{3}$$

In Equation (3), head represents the number of attention heads. $\mathbf{X}\mathbf{W}_i^Q$, $\mathbf{X}\mathbf{W}_i^K$, and $\mathbf{X}\mathbf{W}_i^V$ represent the projection matrices of the i th head query vector, head key vector, and head value vector, $\mathbf{W}^O$ represents the output fusion matrix, and $\text{Concat}()$ represents the concatenation operation, The number of attention heads is 8, which is determined by GPU video memory optimization test. In the process of extracting semantic graph features from English compositions, in order to dynamically capture the relationships within sentences, a semantic graph adjacency matrix is constructed, and its construction formula is shown in Equation (4).

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{E}\mathbf{E}^T}{\sqrt{D}}\right) \tag{4}$$

In Equation (4), $\mathbf{A}$ represents the adjacency matrix, $\mathbf{E}^T$ represents the transpose matrix of the word embedding matrix $\mathbf{E}$, and D is the embedding dimension. Continuing with the study of iteratively updating node features to capture higher-order relationships in semantic graphs, the graph convolution feature propagation formula is shown in Equation (5).

$$\mathbf{H}^{(l+1)} = \sigma\left(\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{\Theta}^{(l)}\right) \tag{5}$$

In Equation (5), $\mathbf{H}^{(l)}$ represents the node feature matrix of the l th layer, $\mathbf{D}^{-\frac{1}{2}}$ is used for normalization, $\mathbf{\Theta}^{(l)}$ represents the learnable weight matrix, σ is the activation function, and introduces nonlinearity to enhance the model's expressive power. Finally, the study integrates all node features and aggregates them into a graph level feature vector to represent the global semantics of the entire English composition. The graph level feature aggregation formula is shown in Equation (6).

$$\begin{cases} \mathbf{z} = \sum_{i=1}^{N} \alpha_i \mathbf{h}_i^{(L)} \\ \alpha_i = \dfrac{\exp(\mathbf{w}^{\cdot}\mathbf{h}_i^{(L)})}{\sum_j \exp(\mathbf{w}^{\cdot}\mathbf{h}_j^{(L)})} \end{cases} \tag{6}$$

In Equation (6), $\mathbf{h}_i^{(L)}$ is the feature vector of the i th node, L is the total number of layers, $\mathbf{z}$ is the graph level feature vector representing the semantic summary of the entire text, α_i is the attention weight representing the importance of node i to global features, $\mathbf{w}$ is the learnable weight vector used to calculate attention scores, and N is the number of nodes or words. In summary, the detection model structure that integrates the sequence features of English compositions with the semantic map features of English compositions is shown in Figure 2.

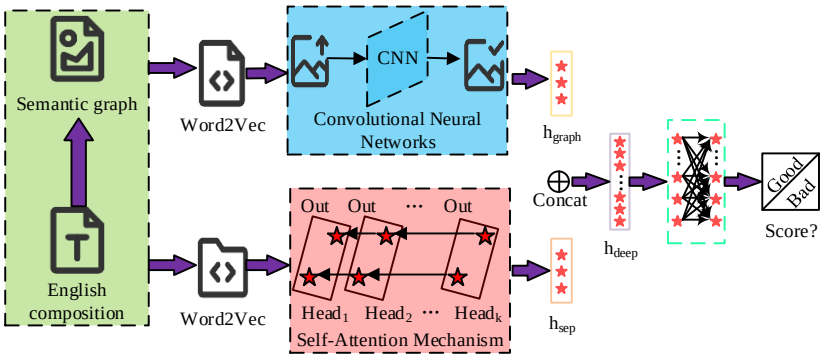


Figure 2: Detection model integrating sequence and graph features of english compositions

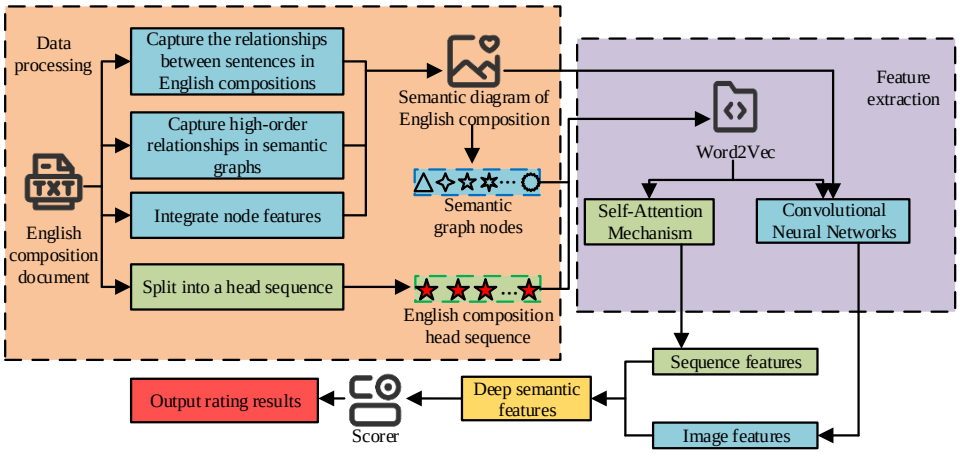


Figure 3: Intelligent scoring model for english compositions based on deep semantic text features

As shown in Figure 2, the detection model that integrates English composition sequence features and semantic graph features receives two types of input data, and the semantic graph captures the semantic relationships between phrases and concepts. Then the semantic graph and English composition are processed by Word2Vec, converting discrete words into dense, low dimensional real valued vectors. Next, semantic graph features are extracted through graph convolutional networks, while sequence features are extracted by introducing attention mechanisms. Subsequently, feature fusion is performed, and the sequence feature vectors and graph feature vectors extracted from two parallel paths are concatenated to form deep semantic features. Finally, the model evaluates the output rating results. The fusion detection model solves the limitation of single feature representation ability by fusing two complementary feature representations. The deep semantic text feature expression of its fusion model is shown in Equation (7).

$$h_{\text{deep}} = \sigma\left(h_{seq} \| h_{graph}\right) \tag{7}$$

In Equation (7), h_{deep} represents the deep semantic features of English composition, $\|$ represented by vector concatenation symbols, h_{seq} and h_{graph} respectively represent the sequence features and graph features of English composition. In summary, the intelligent scoring model for English composition based on deep semantic text features is shown in Figure 3.

As shown in Figure 3, the English composition scoring model based on deep semantic text features achieves accurate evaluation by integrating multi-level semantic information. The model first performs structured parsing on the input English essay document, breaks down the title sequence to highlight the article structure, and preserves contextual information through node feature integration. In the feature extraction stage, multimodal technology is used to deeply fuse semantic information. On the one hand, the title sequence is embedded in Word2Vec and local sequence features are extracted through self-attention mechanism. On the other hand, semantic graph nodes model global semantic relationships through graph convolutional networks. The two types of features are further combined with image features to form a unified deep semantic feature vector. Finally, the rater performs regression analysis based on deep semantic features and outputs objective scoring results.

In summary, the implementation details of the entire research framework are as follows: (1) Word2Vec is used to convert English essay texts into dense word vector matrices. The continuous bag-of-words model predicts core words through contextual word prediction. The input layer aggregates multiple contextual word vectors, while the mapping layer summarizes them to output core word probabilities. The skip-word model predicts contextual words based on core words. Both models undergo

negative sampling optimization, with 300-dimensional embeddings and a 5-window contextual size. (2) During attention mechanism feature extraction, the input word vector matrix is linearly transformed to generate query matrices, key matrices, and value matrices, each with 64 dimensions. The multi-head architecture employs 8 heads, where each head independently computes attention and outputs concatenated linear fusion results. The final sequence features are generated. (3) In graph convolutional network semantic feature extraction, the adjacency matrix embedding dimension is 300. The feature propagation and aggregation process learn a weight matrix with 128 dimensions across 2 layers.

3.2 Intelligent english composition scoring model combined with artificial rules

Although the English composition grading model based on deep semantic text features can effectively grade English compositions, it generally relies on manually defined grading templates, and candidates can avoid deduction types through simple writing techniques, lacking interpretability [19]. In the field of composition checking, artificial rules are usually expressed in formal language and automatically detected through natural language processing tools. The English scoring model combined with artificial rules can effectively solve the problem of lack of interpretability in DL models, so further research is needed to introduce artificial rules [20]. Based on manual rules to quantify the basic language quality of sentences, the basic formula for scoring errors in English compositions is shown in Equation (8).

$$E_s = \frac{1}{1+\lambda\cdot(C_{spell}+C_{gram})}\times F \tag{8}$$

In Equation (8), E_s is the score for incorrect sentences, with a maximum score of F, C_{spell} is the number of spelling errors, C_{gram} is the number of grammar errors, and λ is the error penalty coefficient, its value is 0.1, and the error rate is lowest when its value is 1 through grid search verification. Continuing with the study of balancing the importance of each dimension through artificial rules, the formula for weighting the multidimensional excellence of sentences is shown in Equation (9).

$$Q_s = \beta_1\cdot V_s + \beta_2\cdot G_s + \beta_3\cdot T_s + \beta_4\cdot P_s \tag{9}$$

In Equation (9), Q_s represents the overall excellence score of the sentence, V_s represents the vocabulary score, G_s represents the syntactic complexity score, T_s represents the part of speech diversity score, P_s represents the rectangle score, and β_i represents the artificial rule weight. Then, to evaluate the logical rigor of English essay paragraphs, a scoring formula for paragraph cohesion strength is introduced, and its specific expression is shown in Equation (10).

$$C_p = \frac{\sum_{k=1}^{n}\gamma_k\cdot I(\text{conn}_k)}{\text{N}}\times\frac{R_{\text{cohere}}}{1+\log(\text{L})} \tag{10}$$

In Equation (10), C_p represents the coherence score of the paragraph, $I(\text{conn}_k)$ represents the validity indicator function of the k th connector, γ_k represents the weight of the connector, R_{cohere} represents the semantic coherence ratio, N represents the number of sentences, and L represents the length of the paragraph. Finally, the study aims to achieve dynamic comprehensive scoring of the entire English composition through multi-dimensional manual rule evaluation. The scoring formula is shown in Equation (11)

$$Score = \lambda\cdot\left(\frac{\sum_{j=1}^{m}Q_{s_j}}{m}\right)+\mu\cdot\left(\frac{\sum_{p=1}^{l}C_{p_p}}{l}\right)+\nu\cdot\text{Sim}_{\text{content}}\cdot\text{Sim}_{\text{str}} \tag{11}$$

In Equation (11), $Score$ represents the final score of the composition, Q_{s_j} represents the excellence score of the j th sentence, C_{p_p} represents the coherence score of the p th paragraph, $\text{Sim}_{\text{content}}$ and Sim_{str} represents the similarity of content and structure, λ,μ,ν all meet the requirement of $\lambda+\mu+\nu=1$. The artificial rules are constructed based on expert knowledge, employing a method that quantifies sentence-level errors and sentence excellence through predefined weights to achieve digital transformation. The primary linguistic features targeted include surface errors, sentence-level errors, and paragraph-level errors. By integrating deep semantic features through a Wide&Deep architecture, the rules enhance interpretability while capturing subtle errors and reducing subjective variations. Experimental validation demonstrates their effectiveness in lowering bias values and misjudgment rates, as well as improving scoring stability. In summary, the feature extraction framework for the manual scoring rules of English compositions is shown in Figure 4.

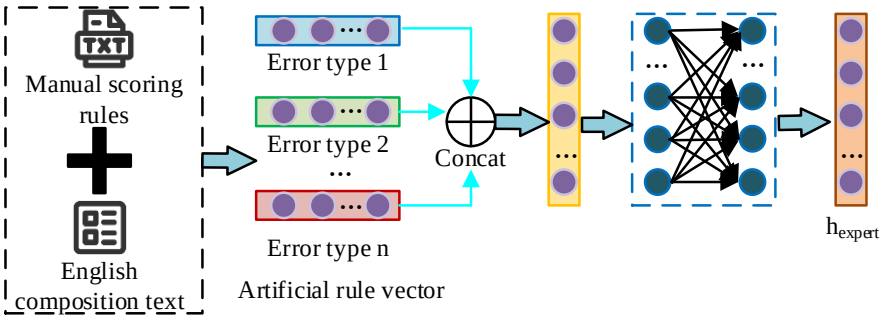


Figure 4: Feature extraction of artificial rules for english composition

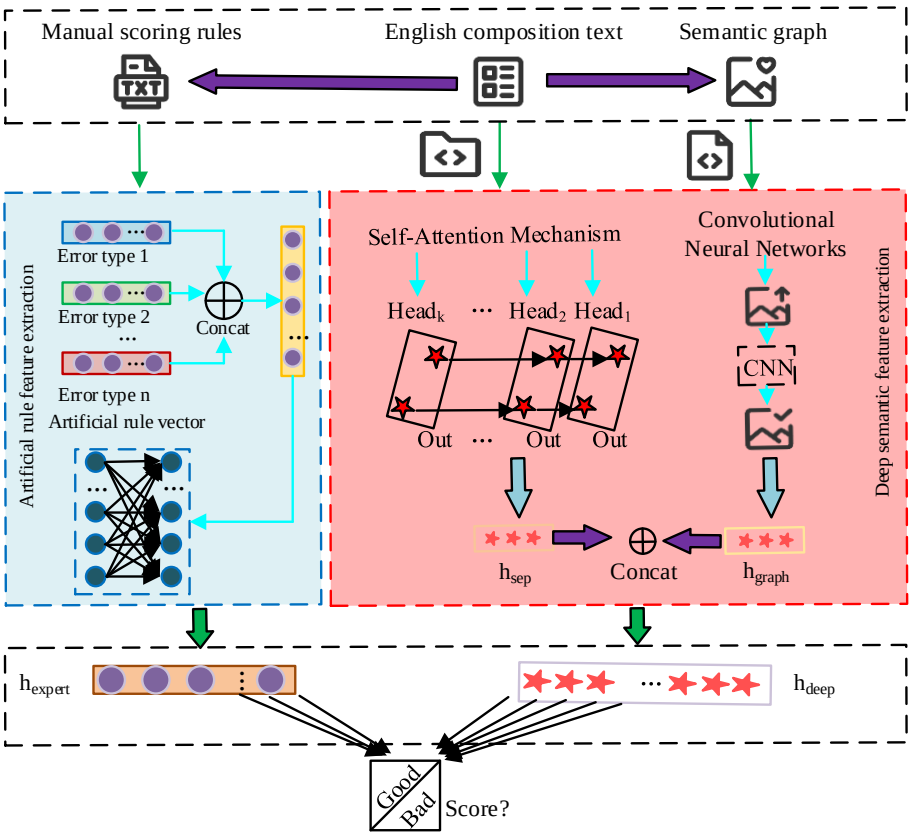


Figure 5: Network structure of english composition error detection combined with artificial rules

As shown in Figure 4, in the feature extraction framework of manual scoring rules for English compositions, structured manual scoring rules are input together with the original English composition text as initial data. Then, the manual rules are decomposed into different types of errors, and each type of rule is quantified as a numerical vector to achieve the digital transformation of expert knowledge. Then, the artificial rule vector is concatenated with the semantic vector of the composition text to form a mixed feature that combines both artificial rules and text semantics. Finally, after processing, output the characteristics of the manual scoring rules for English compositions. The study aims to achieve the organic integration of artificial rules and DL models by converting discrete artificial rules into continuous features. The specific expression is shown in Equation (12).

$$h_{expert} = \sigma(\mathbf{W}(\|_{v \in rud} \mathbf{x}_i) + b) \tag{12}$$

In Equation (12), h_{expert} represents the artificial rule feature, $v \in rud$ represents the set of error types, $\mathbf{x}_i$ represents the artificial rule vector for the error type, and b represents the bias term. Next, the study uses the Wide&Deep structure to fuse shallow features of artificial rules with deep semantic text features, achieving the final error classification prediction. The fusion formula is shown in Equation (13).

$$y = Softmax(\mathbf{W}_{wide} h_{expert} + \mathbf{W}_{deep} h_{deep} + b) \tag{13}$$

In Equation (13), y represents the rating result, and $\mathbf{W}_{wide}$ and $\mathbf{W}_{deep}$ represent the weight matrices of

parts *wide* and *deep*. In summary, the network structure of English composition error detection combined with manual rules is shown in Figure 5.

As shown in Figure 5, the English composition error detection network structure combined with manual rules improves detection accuracy and interpretability through dual channel feature fusion. The model receives dual source inputs: the manual scoring rules are decomposed and vectorized into quantifiable rule vectors, covering error types such as grammar, logic, rhetoric, etc. The model synchronizes the construction of semantic maps for original English compositions, extracts logical relationships between sentences, and performs sequence analysis to capture word order features. Then, the deep semantic features obtained from the dual source input are evaluated together with the artificial rule features, and the results are judged. The study introduces binary cross entropy loss to measure the difference between misclassified predictions and true labels. The specific expression of the loss function is shown in Equation (14).

$$Loss = -\hat{y}\log(y) - (1-\hat{y})\log(1-y) \tag{14}$$

In Equation (14), $Loss$ represents the loss function value, $\hat{y}$ represents the true label of the sample, and y represents the predicted probability output of the model. Finally, the study evaluates the performance of the model by calculating its accuracy, as shown in Equation (15).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{15}$$

In Equation (15), $Accuracy$ represents the accuracy of the model, TP and TN are the number of essays correctly rated as low or high by the model, FP and FN are the number of essays incorrectly rated as low or high by the model. In summary, the scoring process of the English composition scoring model based on DL and artificial rules is shown in Figure 6.

As shown in Figure 6, the English essay scoring model based on DL and artificial rules improves scoring accuracy and interpretability through dual channel feature collaboration. The model takes English composition text and manual scoring rules as dual source inputs: on the one hand, it generates a semantic map through multi-level parsing of the original text, and on the other hand, it breaks down the document into sequential features according to its structure, preserving the framework information of the article. Next, in the feature extraction stage, a bimodal DL architecture is adopted. After Word2Vec vectorization of semantic graph nodes, a graph convolutional network models global semantic relationships and outputs deep features. Sequence nodes extract local language patterns through self-attention mechanisms, generate sequence features, and concatenate the two to form deep semantic features. On the other hand, breaking down manual rules in textual form into quantifiable dimensional vectors enables the digital transformation of expert knowledge. Finally, the artificial rule features are optimized using binary cross entropy and combined with deep semantic features to generate rule enhanced deep features. The rater then performs regression analysis based on the rule enhanced deep features to output the final English essay grading results.

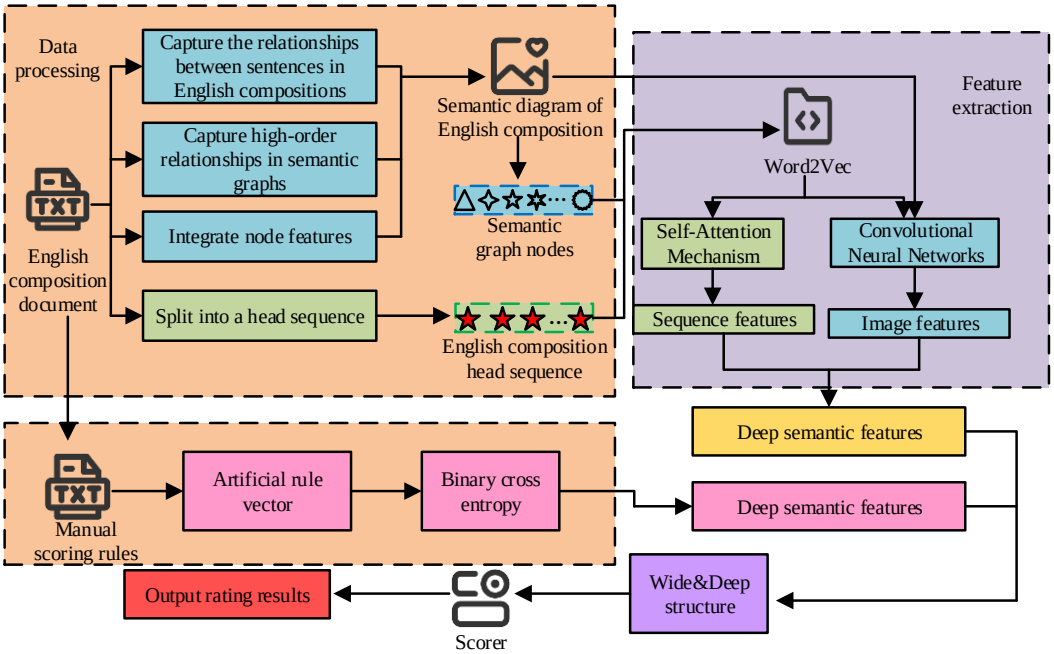


Figure 6: Scoring process of english composition scoring model based on dl and artificial rules

4 Validation of English composition grading model based on DL and artificial rules

4.1 Performance testing of English composition scoring model based on DL and artificial rules

To confirm the capability of the English essay grading model based on DL and artificial rules, a simulation model was constructed for testing. The testing environment and specific configuration are presented in Table 1.

Table 1: Test environment and specific configuration

Testing environment	Specific configuration
GPU	NVIDIA Tesla V100/A100
CPU	Intel Xeon Gold 6248R
Memory	256GB DDR4
Storage	2TB NVMe SSD + 10TB HDD
DL framework	PyTorch 1.12 / TensorFlow 2.10
Feature engineering tools	Scikit-learn 1.2 + Gensim 4.3
Support for large models	Transformers 4.28

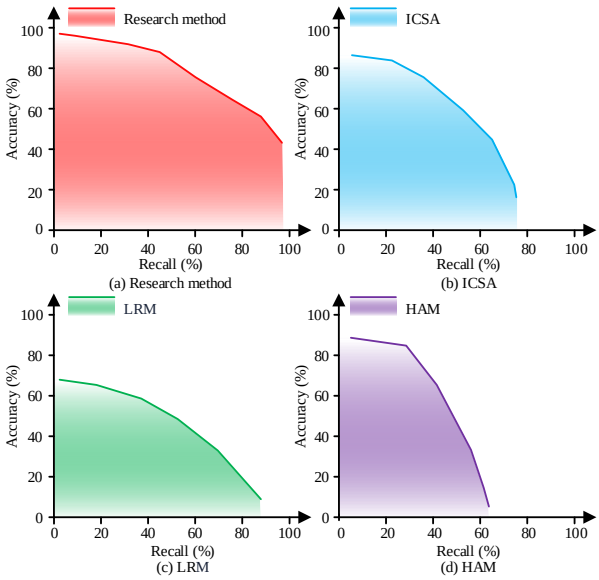


Figure 7: Accuracy recall curve of different methods

As shown in Table 1, the specific configurations in the table were used for performance testing, using the Kaggle ASAP dataset. The research methods were compared with the Integrated Classification Scoring Algorithm (ICSA), Linear Regression Model (LRM), and Hierarchical Attention Model (HAM). The accuracy recall curves and curve areas of the four methods were compared, and the results are presented in Figure 7.

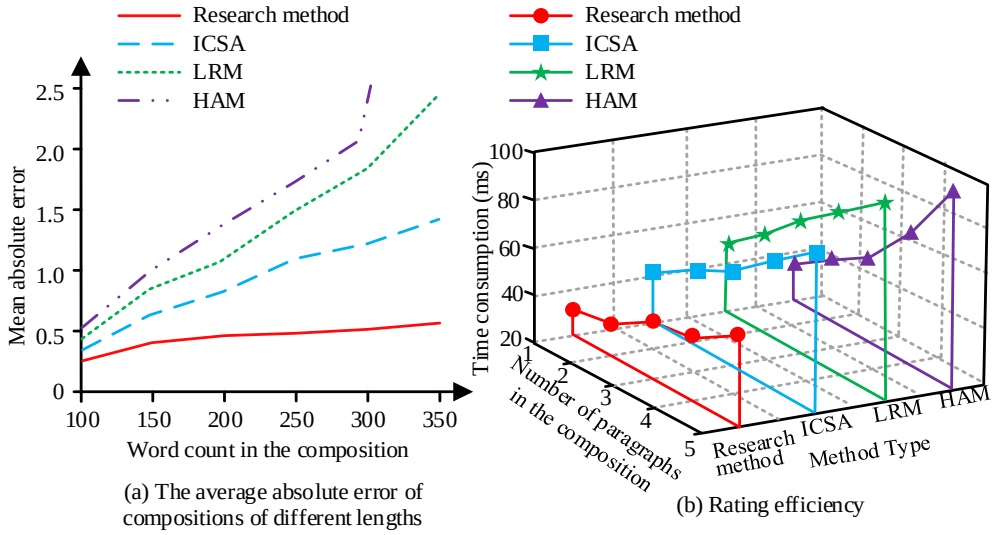
As shown in Figure 7, the shape and area of the accuracy recall curve of different methods are different. In Figure 7 (a), the accuracy recall curve of the research method was close to a rectangle, with a curve area of 92.3%. In Figure 7 (b), the accuracy recall curve of the ICSA algorithm was 71.6%. In Figure 7 (c), the curve of the LRM model belonged to low accuracy high recall, which was prone to false positives. As shown in Figure 7 (d), the curve of the HAM model belonged to high accuracy low recall, which was prone to missed detections. Overall, compared to comparative methods, research methods had higher accuracy and inspection coverage. The mean absolute error (MAE) of the scoring results of the four methods under different numbers of writing words, as well as the scoring time under different numbers of writing paragraphs, were compared, and the outcomes are presented in Figure 8.

In Figure 8 (a), the MAEs of the scoring results of the four methods all increased with the increase in the number of English composition words. The MAE of the research method's scoring results had the smallest increase. When the word count in the composition was 100, the MAE of the research method's scoring results was 0.25. When the word count in the composition was 350, the MAE of the research method's scoring results was 0.52. The MAE for the two types of composition word counts only increased by 0.27. The MAE of the scoring results for the other three methods was significantly greater than that of the research method at different numbers of words in the composition. In Figure 8 (b), the scoring time of all four methods increased with the number of paragraphs in the essay. When the English essay had only one paragraph, the scoring time of the research method was 32 ms, and when the essay had five paragraphs, the scoring time was 42 ms. However, the scoring time of the other three methods at different paragraph counts was significantly greater than that of the research method. Overall, compared to the comparative methods, the research methods had better robustness. In conclusion, the English essay grading model proposed by the research based on DL and artificial rules had high reliability, accuracy, and good robustness. After validating the performance of the research methodology, the study further investigated the synergistic effects of the fusion architecture through ablation experiments. First, independent testing of deep models revealed that removing manual rules reduced grammatical error detection accuracy, demonstrating their constraint effect on surface errors. Next, independent testing of rule models showed increased semantic coherence score deviations in long texts when graph convolutional networks were removed, proving deep models' capability to capture higher-order semantics. Finally, dual-stream feature contribution analysis using SHAP values demonstrated that manual rule features contributed minimally to grammatical/spelling error detection, while

deep semantic features played a significant role in content logic scoring. These ablation results confirmed the complementary innovation of the "feature perception-regulation constraint" architecture in the research methodology.

4.2 Practical application effect of English composition scoring model based on DL and artificial rules

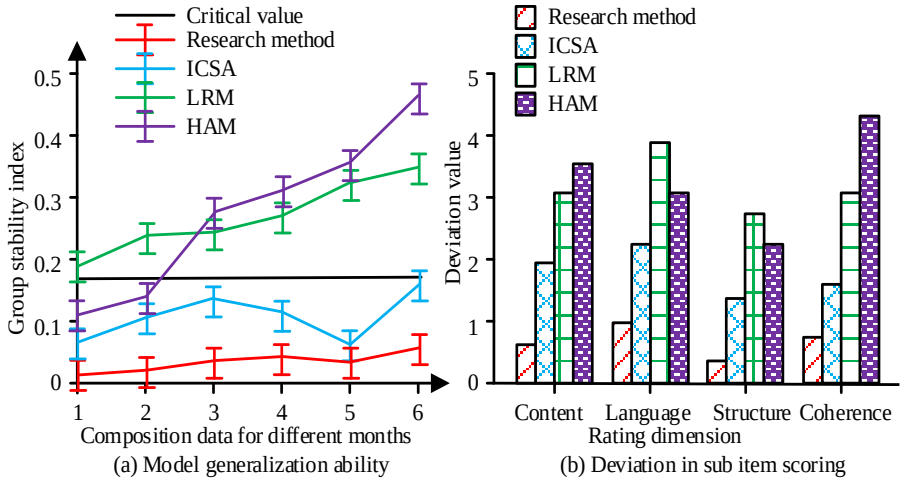
On the basis of verifying the performance of the English essay grading model based on DL and artificial rules, further research is conducted to ascertain the efficacy of the practical application of the research method. The study used the IELTS Writing Task 2 dataset to build a modular hierarchical architecture experimental platform. The research methods were compared with ICSA, LRM and HAM, and the semantic depth index was supplemented: The content dimension deviation of BERT in IELTS data set could be reduced to 0.42, which was better than the 0.67 of the research models, highlighting the advantages of lightweight. The study further compared Transformer-based pre-trained models. In the IELTS content dimension scoring, BERT exhibited lower deviation values than the research methodology. However, its reliance on billions of parameters resulted in significantly longer response times. Notably, the research methodology demonstrated markedly higher accuracy in detecting grammatical errors when incorporating rules, surpassing BERT. These findings indicated that compared to cutting-edge technologies, the research methodology demonstrated superior semantic understanding depth and error detection specificity. Then, the group stability index of the four methods for the English composition data in the first six months was scored, and the deviation values under different scoring dimensions were compared. The results are shown in Figure 9.



(a) The average absolute error of compositions of different lengths

(b) Rating efficiency

Figure 8: MAE and rating efficiency



(a) Model generalization ability

(b) Deviation in sub item scoring

Figure 9: Model generalization ability and sub-item scoring deviation

In Figure 9 (a), the critical value of the group stability index for English composition scoring was 0.17. The overall group stability index of the research method for scoring monthly composition data remained below

the critical value, with its highest group stability index being 0.07 in June and the lowest 0.02 in January. The stability indices of the other three methods for scoring monthly essay data were significantly higher than that of the research method. In Figure 9 (b), the deviation value of the research method in the dimension of composition content was 0.67, the deviation value in the dimension of composition language was 0.99, the deviation value in the dimension of composition structure was 0.33, and the deviation value in the dimension of composition coherence was 0.82. The bias values of the other three methods under different scoring dimensions were significantly greater than those of the research methods. Overall, compared to the comparative methods, the research methods had better generalization ability and higher accuracy. Comparing the sensitivity of four methods in identifying excellent compositions and their ability to capture advanced vocabulary in compositions, the results are presented in Figure 10.

As shown in Figure 10 (a), the misjudgment rates of the four methods for high - scoring essays with different score thresholds were not the same. The overall misjudgment rate of the research method for high-scoring essays was less than 20%. The highest misjudgment rate was 17.8% when the score threshold was 21 points, and the minimum misjudgment rate of the research method was 3.2% when the score threshold was 24-25 points. The misjudgment rates of the other three methods were significantly higher than that of the research method at different score thresholds. As shown in Figure 10 (b), the consistency between the vocabulary scoring results of the four methods and the manual vocabulary scoring was not the same. The distribution of the vocabulary scoring results of the research method was closely aligned with the diagonal, indicating it was highly consistent with the manual vocabulary scoring. However, the distribution of vocabulary scoring results for the other three methods differed significantly from that of manual vocabulary scoring, resulting in lower accuracy of their scoring results. Overall, compared to the comparative methods, the research method had a lower false positive rate and better scoring performance. The four methods were compared for the accuracy rate of scoring under different error types and the average response time under different concurrent request numbers, as shown in Figure 11.

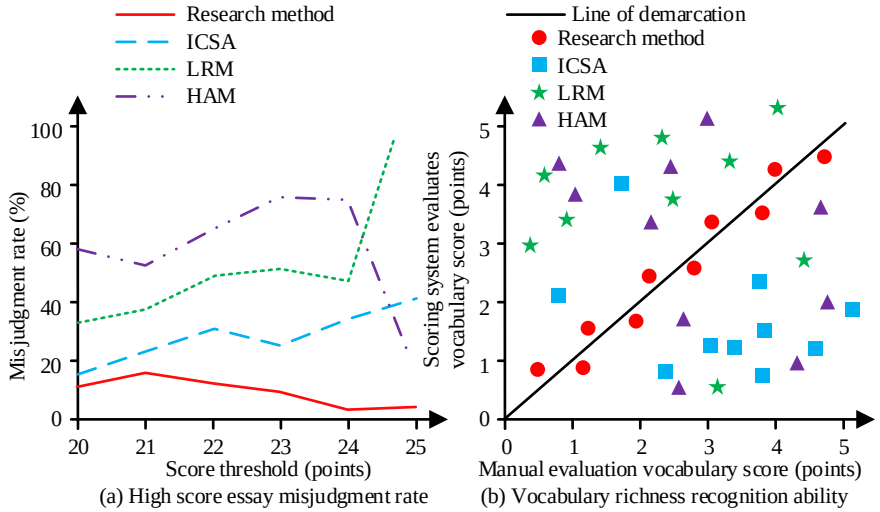


Figure 10: High score essay misjudgment rate and vocabulary richness recognition ability

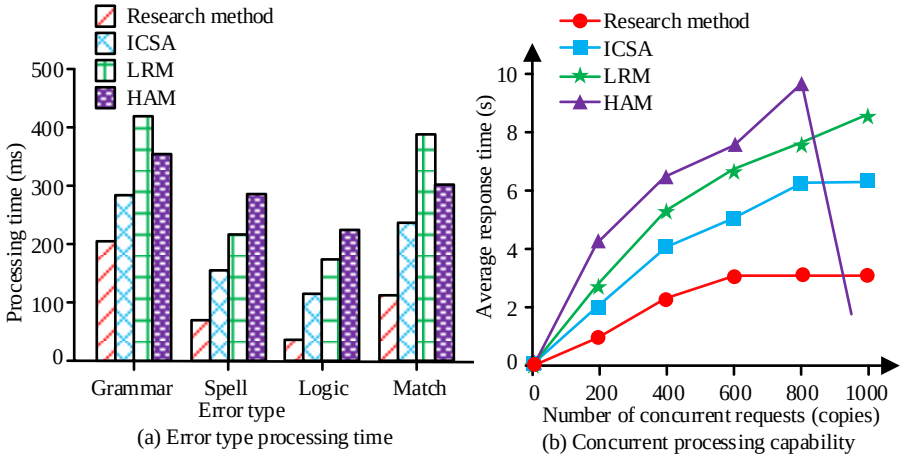


Figure 11: Error type processing time and concurrent processing capability

As shown in Figure 11(a), the research method demonstrated 99.2% accuracy in scoring grammatical errors, 98.7% in spelling errors, 99.0% in logical errors, and 97.9% in collocation errors. In contrast, the other three methods showed significantly lower accuracy rates for these error types compared to the research methodology. In Figure 11 (b), the average response time of the four methods gradually increased with the increase of concurrent requests, with the research method showing the smoothest trend of increase. When faced with 600 concurrent requests, the average response time of the research method reached a stable value of 3.4 seconds. However, the average response time of the other three methods showed a significantly greater increase trend than the research method. Overall, compared to comparative methods, research methods had better resource allocation capabilities and scoring performance. Overall, the English essay grading model proposed by the research based on DL and artificial rules had good generalization ability, accuracy, and performance.

5 Conclusion

To address the issues of high misjudgment rates and instability in existing English essay automatic scoring systems, this study innovatively proposes an English essay scoring model combining DL with manual rules. The research methodology extracts sequence features and semantic graph features from English essays, integrating them with manual rule features to construct a "feature perception-rule constraint-joint decision" fusion architecture for stable and accurate scoring. Experimental results show that when the essay contains 100 words, the average absolute error of the scoring method is 0.25; when the essay contains 350 words, the average absolute error increases to 0.52; and when the essay consists of 5 paragraphs, the scoring time reaches 42ms. In practical application tests, the method shows 0.67 deviation in content dimension scoring, 0.99 deviation in language dimension scoring, 0.33 deviation in structure dimension scoring, and 0.82 deviation in coherence dimension scoring. The method achieved 99.2% accuracy rate for grammatical errors, 98.7% accuracy rate for spelling errors, 99.0% accuracy rate for logical errors, and 97.9% accuracy rate for collocation errors. Overall, the proposed method demonstrated excellent scoring accuracy, robustness, and stability. The research findings failed to quantify the contribution ratios of DL and rule-based approaches to explainability. The test datasets were limited to IELTS/Kaggle materials, which did not validate the generalization capabilities of open-domain essays and consequently compromised practical applicability. Moreover, the methodology primarily relied on Word2Vec and traditional attention mechanisms for feature extraction. While effective in English essay scoring, the static embedding model of Word2Vec lacked contextual sensitivity, potentially limiting semantic depth comprehension and cross-linguistic transfer capabilities. Modern Transformer models, however, provide superior contextual representation and enhanced cross-linguistic application potential. Future studies could integrate Transformer pre-trained models to verify model stability and deviations across multilingual essay datasets (e.g., French, Chinese), evaluate cross-linguistic rule adaptability, and improve cross-linguistic performance and transferability. Additionally, the research could incorporate eye-tracking technology into multi-modal deep understanding frameworks. By recording eye movements during writing processes, it can analyze authors' attention allocation patterns. Combined with keystroke logs, this approach could quantify writing fluency and cognitive load, supplementing process dynamics that textual features cannot capture. However, this study is the first to migrate the Wide&Deep architecture from recommendation system to essay scoring field. Through semantic drift of DL with rule feature constraints, it provides a new idea for the interpretability of AI education products.

References

[1] Del Gobbo E, Guarino A, Cafarelli B, Grilli L. GradeAid: A framework for automatic short answers grading in educational contexts—design, implementation and evaluation. Knowledge and Information Systems, 2023, 65(10): 4295-4334. DOI: 10.1007/s10115-023-01892-9.

[2] Wang Q. The use of semantic similarity tools in automated content scoring of fact-based essays written by EFL learners. Education and Information Technologies, 2022, 27(9): 13021-13049. DOI: 10.1007/s10639-022-11179-1.

[3] Geçkin V, Kızıltaş E, Çınar Ç. Assessing second-language academic writing: AI vs. Human raters. Journal of Educational Technology and Online Learning, 2023, 6(4): 1096-1108. DOI: 10.31681/jetol.1336599.

[4] Theodosiou A A, Read R C. Artificial intelligence, machine learning and deep learning: Potential resources for the infection clinician. Journal of Infection, 2023, 87(4): 287-294. DOI: 10.1016/j. Jinf.2023.07.006.

[5] Wang J, Wang S, Zhang Y. Deep learning on medical image analysis. CAAI Transactions on Intelligence Technology, 2025, 10(1): 1-35. DOI:10.1049/cit2.12356.

[6] Ramesh D, Sanampudi S K. An automated essay scoring systems: A systematic literature review. Artificial Intelligence Review, 2022, 55(3): 2495-2527. DOI: 10.1007/s10462-021-10068-2.

[7] Fokides E, Peristeraki E. Comparing ChatGPTs correction and feedback comments with that of educators in the context of primary students short essays written in English and Greek. Education and Information Technologies, 2025, 30(2): 2577-2621. DOI: 10.1007/s10639-024-12912-8.

[8] Shahzad A, Wali A. Computerization of off-topic essay detection: a possibility? Education and Information Technologies, 2022, 27(4): 5737-5747. DOI: 10.1007/s10639-021-10863-y.

[9] Erturk S, van Tilburg W A P, Igou E R. Off the mark: Repetitive marking undermines essay evaluations due to boredom. Motivation and Emotion, 2022, 46(2): 264-275. DOI: 10.1007/s11031-022-09929-2.

[10] Sharma A, Katlaa R, Kaur G, Jayagopi D B. Full-page handwriting recognition and automated essay scoring for in-the-wild essays. Multimedia Tools and Applications, 2023, 82(23): 35253-35276. DOI: 10.1007/s11042-023-14558-z.

[11] Mohammed A, Kora R. A comprehensive review on ensemble deep learning: Opportunities and challenges. Journal of King Saud University-Computer and Information Sciences, 2023, 35(2): 757-774. DOI: 10.1016/j. Jksuci.2023.01.014.

[12] Tropsha A, Isayev O, Varnek A, Schneider G, Cherkasov A. Integrating QSAR modelling and deep learning in drug discovery: The emergence of deep QSAR. Nature Reviews Drug Discovery, 2024, 23(2): 141-155. DOI: 10.1038/s41573-023-00832-0

[13] Whang S E, Roh Y, Song H, Lee J G. Data collection and quality challenges in deep learning: A data-centric ai perspective. The VLDB Journal, 2023, 32(4): 791-813. DOI: 10.1007/s00778-022-00775-9.

[14] Pereira T D, Tabris N, Matsliah A, Turner D M, Li J, Ravindranath S, et al. SLEAP: A deep learning system for multi-animal pose tracking. Nature methods, 2022, 19(4): 486-495. DOI: 10.1038/s41592-022-01426-1.

[15] Olan F, Spanaki K, Ahmed W, Zhao G. Enabling explainable artificial intelligence capabilities in supply chain decision support making. Production Planning & Control, 2025, 36(6): 808-819. DOI:10.1080/09537287.2024.2313514.

[16] Bhat M, Rabindranath M, Chara B S, Simonetto D A. Artificial intelligence, machine learning, and deep learning in liver transplantation. Journal of hepatology, 2023, 78(6): 1216-1233. DOI: 10.1016/j. Jhep.2023.01.006.

[17] Simon K, Vicent M, Addah K, Bamutura D, Atwiine B, Nanjebe D, Mukama A O. Comparison of deep learning techniques in detection of sickle cell disease. AIA, 2023, 1(4):252-259. DOI: https://doi.Org/10.47852/bonviewAIA3202853.

[18] K. Bhosle, V. Musande. Evaluation of deep learning CNN Model for recognition of Devanagari digit. Applied Artificial Intelligence. 2023, 1(2): 114-118. DOI: 10.47852/bonviewAIA3202441.

[19] Zamfiroiu A, Vasile D, Savu D. ChatGPT–a systematic review of published research papers. Informatica Economica, 2023, 27(1): 5-16. DOI: 10.24818/issn14531305/27.1.2023.01.

[20] Didimo W, Grilli L, Liotta G, Montecchiani F. Efficient and trustworthy decision making through human-in-the-loop visual analytics: A case study on tax risk assessment. Rivista italiana di informatica e diritto, 2022, 4(2): 15-21. DOI: 10.32091/RIID0092.