Proceedings of the 3rd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2025.27800

Quality Prediction of RAG System Retrieval Based on Machine Learning Algorithms

Chaoyi Yu

School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China
2387663175@qq.com

Abstract. Retrieval Enhanced Generation (RAG) system improves the accuracy and reliability of content Generation by retrieving external knowledge, and has been widely used in intelligent question answering, knowledge assistant and other fields. However, its core performance depends on the quality of the retrieval stage, and the relevance and factual consistency of the retrieval results directly determine the validity of the generated content. However, factors such as query complexity, document noise, and domain differences in real-world scenarios can easily lead to fluctuations in retrieval quality. Traditional manual evaluation is costly and outdated, and it is difficult to meet the real-time optimization requirements. At the same time, the existing models have limitations in complex feature fusion and parameter optimization. Therefore, this paper proposes a retrieval quality prediction model combining lizard optimization algorithm (HLOA), convolutional neural network (CNN), and bidirectional gated recursive unit (BIGRU). The correlation analysis shows that there is a strong positive correlation between the search level and the search usefulness score, that is, the higher the search level, the better the search usefulness score; There is a strong negative correlation between query complexity and retrieval usefulness score, which means that the higher the query complexity, the lower the retrieval usefulness score. Comparing this model with 9 models including decision tree, random forest Adaboost, gradient boosting tree, ExtraTrees, CatBoost, XGBoost, LightGBM, and KNN, it is shown that their performance is better overall: MSE (28.617), RMSE (5.349), MAE (4.401), and MAPE (17.355) are the lowest, while R^2 (0.952) is the highest. This study provides an effective solution for accurate prediction and real-time optimization of RAG system retrieval quality, which helps to enhance the application value of RAG technology in practical scenarios.

Keywords: RAG, Horn lizard optimization algorithm, integrated neural network, bidirectional gated recurrent unit, retrieval quality

1. Introduction

Retrieval enhanced generation (RAG) system improves the accuracy and reliability of generated content by retrieving external knowledge, and has been widely used in intelligent question answering, knowledge assistant and other fields. Its core performance depends on the quality of the

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.27800

retrieval phase-the relevance and factual consistency of the retrieval results directly determine the effectiveness of the generated content. However, in real-world scenarios, factors such as query complexity, document noise, and domain differences usually lead to fluctuations in retrieval quality. Traditional manual evaluation is costly and outdated, and it is difficult to meet the real-time optimization requirements [1,2]. Therefore, constructing efficient retrieval quality prediction model becomes the key to the implementation of RAG system. It is necessary to quantify the matching degree of search results and queries, document credibility and other indicators, predict the search effect in advance, and provide a basis for the adjustment in the subsequent generation process.

Machine learning algorithms provide an automated solution for predicting the quality of RAG retrieval [3]. The traditional model realizes the preliminary prediction of retrieval relevance by mining artificial features such as keyword overlap and entity coverage; Deep learning models can automatically extract deep semantic features from text, capture implicit associations between queries and documents, and significantly improve prediction accuracy. CNN can capture local semantic matching patterns, while RNN models can handle sequence dependencies, effectively handling long texts and complex sentence structures. These algorithms reduce the reliance on artificial feature engineering and can dynamically adapt to data from different domains, providing decision support for real-time optimization of RAG systems. However, there are still problems such as incomplete feature extraction and insufficient model parameter optimization [4].

Aiming at the limitations of existing models in complex feature fusion and parameter optimization, this paper proposes a prediction model combining lizard optimization algorithm (HLOA), convolutional neural network (CNN), and bidirectional gated recursive unit (BIGRU). CNN is used to extract local key features between queries and documents, capturing matching patterns at keyword and phrase levels; BIGRU models the bidirectional context dependencies of text sequences, enhancing the capture of remote semantic associations; HLOA optimizes the key parameters of CNN and BIGRU by simulating the global optimization capability of horned lizard predation behavior, avoiding the model falling into local optimum. The cooperation of the three can fully integrate local and global features, improve the fitting ability of complex patterns in the retrieval quality, and provide a new solution for accurate prediction of retrieval quality in RAG system.

2. Data sources

This dataset is mainly used for quality prediction retrieval in RAG systems, consisting of 754 records, 20 independent variables, and 1 predictor variable. The independent variables belong to four types of features: query features involve query complexity, length, ambiguity, type, and domain; The document features include length, number of days to date, authority, format, factual information, noise level, and structural complexity; Matching features include relevance score, semantic similarity, keyword overlap percentage, entity coverage, domain matching degree, and language consistency; The retrieval features include ranking and retrieval methods in the retrieval results. The predictor variable is the retrieval usefulness score. This dataset can be used to train machine learning models to predict the actual level of assistance that retrieval results can provide for generating tasks, providing support for optimizing the performance of RAG systems. Table 1 shows the statistical results of the dataset.

DOI: 10.54254/2755-2721/2025.27800

Table 1. The statistical results of the dataset

Variable Name	Mean	Variance	Median
query_complexity	5.401857	8.578005	5
query_length	16.037135	6.295479	16
query_ambiguity	0.506777	8.523993	0.515
query_type	3.018568	2.012935	3
query_domain	5.404509	8.177456	5
document_length	1029.218833	2.998369	1001.5
document_age_days	1844.591512	1.096582	1826.5
document_authority	0.486671	8.079143	0.485
document_format	2.961538	2.021095	3
document_factuality	0.639416	4.267403	0.63
document_noise	0.352573	4.020825	0.36
document_structural_complexity	2.891247	2.057214	3
relevance_score	0.517308	8.236977	0.53
semantic_similarity	0.502586	8.772836	0.495
keyword_overlap	48.888727	8.065791	50.45
entity_coverage	0.515491	8.090155	0.51
domain_match	0.488992	8.362395	0.48
language_consistency	0.746711	2.087629	0.74
retrieval_rank	10.159151	3.461474	10
retrieval_method	2.48939	1.259515	2.5
retrieval_usefulness_score	41.673873	6.092095	42.1

FIG. 1 shows a related heat map. From these correlation scores, it can be inferred that there is a strong positive correlation between retroval_rank and retroval_usefilitess_store, indicating that search ranking may have a significant positive effect on search usefulness scores. The higher the search rank, the higher the search usefulness score may be [5]. And there is a strong negative correlation between query_completeness and retrieval_usefulness_store, which indicates that the higher the query complexity, the lower the retrieval usefulness score. The correlation between other variables and retrieval_usefilitess_store is relatively weak, and their effect on the retrieval usefulness score is relatively limited. However, they can also reflect to some extent the weak relationship between various factors and the usefulness of retrieval. Document_factual has a certain positive correlation, indicating that the factual nature of the Document may have a positive impact on retrieval usefulness [6].

DOI: 10.54254/2755-2721/2025.27800

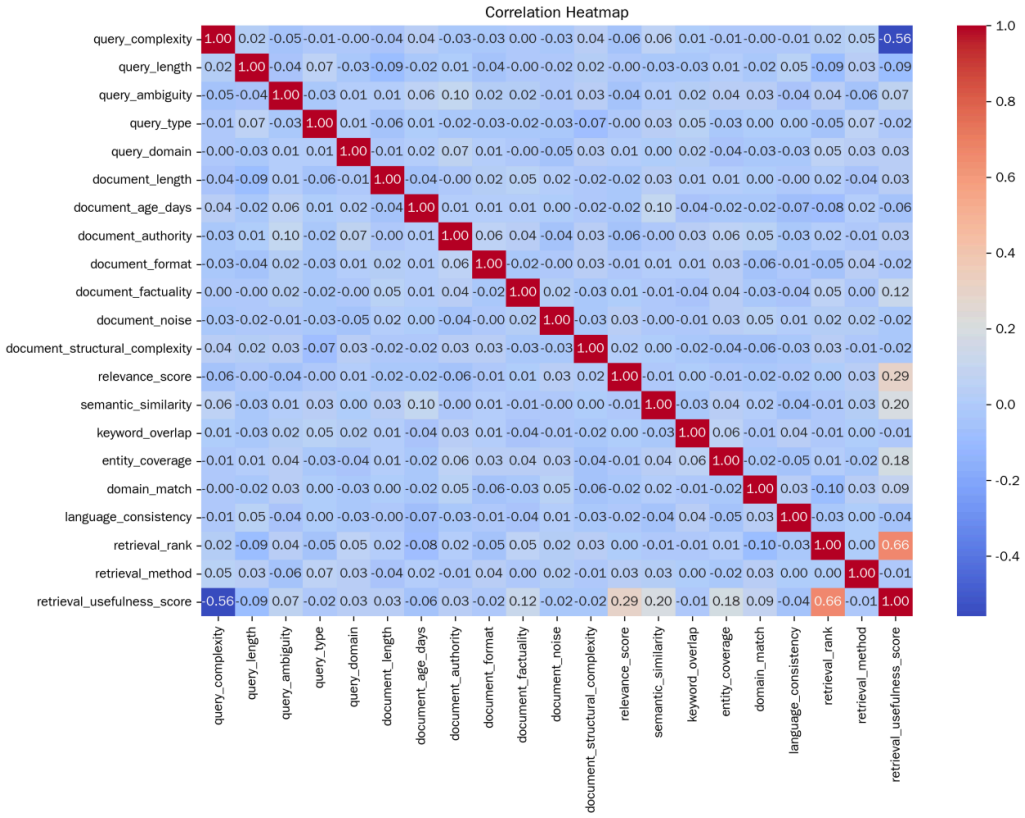


Figure 1. Correlation analysis

Horned lizard optimization algorithm (HLOA) is a meta-heuristic optimization algorithm inspired by the survival behavior of horned lizards. Its core principles simulate efficient hunting and adaptation strategies of horned lizards in desert environments [7]. The network architecture of HLOA is shown in Figure 2. In the initialization stage of the algorithm, the solution space of the problem to be optimized is mapped to a "habitat", and the randomly generated candidate solutions are used as "horned lizard individuals". The search process is divided into two phases: exploration and development. During the exploration phase, the behavior of large-scale foraging in horned lizards was simulated. Carry out a global search in the solution space by randomly adjusting the step size and direction to avoid falling into the local optimum; In the development phase, it simulates the precise sprint of horned lizards after targeting their prey, narrows the search range based on the current optimal solution, and improves the local search accuracy through fine-tuning [8]. In addition, the algorithm also introduces a "temperature regulation" mechanism to adaptively balance the weights of exploration and development according to the iterative process — initially focusing on global exploration to cover a broader solution space, and later enhancing local development to refine the optimal solution. This strategy of simulating biological adaptive behavior makes HLOA have strong global optimization ability and convergence speed in complex optimization problems.

3. Method

3.1. HLOA

The Horned Lizard Optimization Algorithm (HLOA) is a metaheuristic optimization algorithm inspired by the survival behavior of horned lizards. Its core principle simulates the efficient hunting

DOI: 10.54254/2755-2721/2025.27800

and adaptation strategies of horned lizards in desert environments [7]. The network architecture of HLOA is shown in Figure 2.

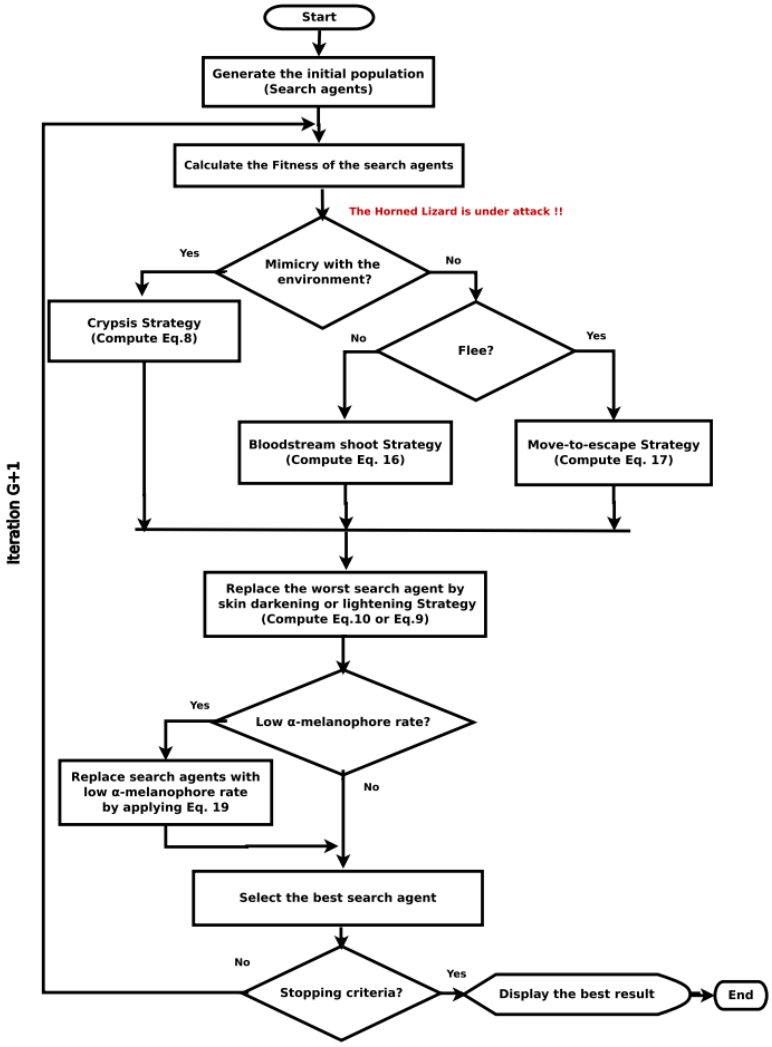


Figure 2. The network architecture of HLOA

In the initialization stage of the algorithm, the solution space of the problem to be optimized is mapped to a "habitat", and the randomly generated candidate solutions are used as "horned lizard individuals". The search process is divided into two phases: exploration and development. During the exploration phase, the behavior of large-scale foraging in horned lizards was simulated. Carry out a global search in the solution space by randomly adjusting the step size and direction to avoid falling into the local optimum; In the development phase, it simulates the precise sprint of horned lizards after targeting their prey, narrows the search range based on the current optimal solution, and improves the local search accuracy through fine-tuning [8]. In addition, the algorithm also introduces a "temperature regulation" mechanism to adaptively balance the weights of exploration and development according to the iterative process — initially focusing on global exploration to cover a broader solution space, and later enhancing local development to refine the optimal solution.

DOI: 10.54254/2755-2721/2025.27800

3.2. CNN

Convolutional neural networks (CNNs) are deep learning models inspired by biological visual systems. Its core components include a convolutional layer, a pooling layer, and a fully connected layer. The convolutional layer performs a local convolution operation on the input data by sliding the convolution kernel to extract local features, and the same convolution kernel shares weights throughout the input space, significantly reducing the number of parameters. The pooling layer compresses the feature map dimensions by downsampling, retaining critical information while enhancing translational invariance [9]. Finally, the advanced features are mapped to the output target through a fully connected layer.

3.3. BIGRU

Bidirectional Gated Recursive Unit (BIGRU) is a bidirectional extension model based on Gated Recursive Unit (GRU) specifically designed for processing sequential data. Its core inherits the gating mechanism of GRU, and dynamically regulates the information flow through the update gate and the reset gate: the update gate determines how much historical information to retain, and the reset gate controls whether to ignore the past state, which effectively alleviates the gradient disappearance problem of traditional RNN and enhances the ability to capture long sequence dependencies. Meanwhile, BIGRU introduces a bidirectional structure, including two GRU units, forward and backward. The forward unit processes the information in sequence, while the backward unit processes the information in the reverse order. These two outputs are connected and used as final features [10]. This design enables the model to leverage past and future contextual information to more accurately understand the bi-directional association of elements in a sequence. The network structure of BIGRU is shown in Figure 3.

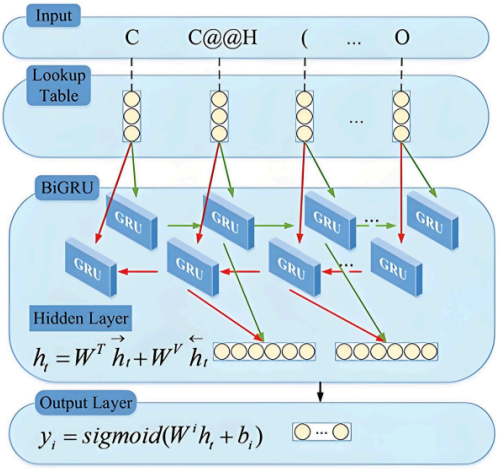


Figure 3. The network structure of BIGRU

3.4. HLOA-CNN-BIGRU

The HLOA and CNN optimized BIGRU model is a hybrid model that combines meta-heuristic optimization and deep learning. In its core architecture, CNN extracts local features through a convolution kernel; BIGRU models contextual dependencies bidirectionally.

HLOA plays a role in parameter optimization in this process: by simulating the exploration development strategy of horned lizard, it adaptively adjusts key parameters such as convolution

DOI: 10.54254/2755-2721/2025.27800

kernel size, stride length, hidden layer dimension of BIGRU, and gate weight of CNN. The cooperation of the three makes the model have the ability of local feature capture, global dependency modeling and parameter optimization efficiency, and is suitable for the prediction task of complex sequence data, especially the retrieval quality evaluation scenario of RAG system.

4. Result

In terms of parameter setting of the model, the number of search agents is 8, the maximum number of iterations is 6, the optimization dimension is 3, the learning rate ranges from 1e-4 to 1e-2, the number of BiGRU neurons ranges from 10 to 30, and the L2 regularization parameter ranges from 1e-4 to 1e-1. Adopting Adam optimization algorithm, the maximum training rounds of the main program are 5 rounds, and the maximum training rounds of the objective function are 500 rounds. The initial learning rate is determined by the optimal value obtained by optimization, and the learning rate scheduling method is segmented. The learning rate decay factor is 0.1, the decay period is 4 in the main program and 400 in the objective function. The L2 regularization parameter is the optimal value obtained by optimization, and the data will be shuffled with each training round. In the network structure, there are two convolutional layers with kernel size [2,1] with 16 and 32 filters, respectively, and the activation function is relu; The number of neurons in the BiGRU layer is the optimal value obtained by optimization, and the output mode is last; It also includes a sequence fold layer, a sequence unfold layer, a flattened layer, a fully connected layer, and a regression layer. In terms of data processing, the training set accounts for 70% of the total data, which is normalized to a range of 0 to 1 using mapminmax.

The comparison models used in this paper include Decision Tree, Random Forest Adaboost, Gradient Boost Tree, Extra Tree, CatBoost, XGBoost, LightGBM, and KNN. The experimental results are shown in Table 2, and the comparison bar chart of each model index is shown in Figure 4.

Table 2. The results of the comparative experiment

Model	MSE	RMSE	MAE	MAPE	R^2
Decision tree	209.006	14.457	11.384	43.368	0.656
Random Forest	105.254	10.259	8.358	26.692	0.803
Adaboost	79.913	8.939	7.189	27.83	0.861
Gradient Boosting Tree (GBDT)	73.064	8.548	6.923	27.185	0.874
ExtraTrees	85.037	9.222	7.534	27.082	0.84
CatBoost	131.626	11.473	9.182	27.276	0.784
XGBoost	49.006	7	5.703	26.013	0.919
LightGBM	114.129	10.683	8.489	32.406	0.809
KNN	52.506	7.246	5.759	25.826	0.909
HLOA-CNN-BIGRU	28.617	5.349	4.401	17.355	0.952

By various metrics, HLOA-CNN-BIGRU outperformed all other models. On MSE, the value of HLOA-CNN-BIGRU is 28.617, much lower than 209.006 for Decision Tree, 105.254 for Random Forest, 79.913 for Adaboost, 73.064 for Gradient Enhancement Tree, 85.037 for Extra Tree, 131.626 for CatBoost, 114.129 for LightGBM, 49.006 for XGBoost, and 52.506 for KNN. In terms of

DOI: 10.54254/2755-2721/2025.27800

RMSE, HLOA-CNN-BIGRU is 5.349, which is lower than 14.457 for Decision Tree, 10.259 for Random Forest, 8.939 for AdaBoosts, 8.548 for Gradient Boosting Tree, 9.222 for ExtraTrees, 11.473 for CatBoosts, 10.683 for LightGBM, 7 for XGBoosts, and 7.246 for KNN. In the MAE index, HLOA-CNN-BIGRU is 4.401, which is lower than 11.384 for Decision Tree, 8.358 for Random Forest, 7.189 for AdaBoosts, 6.923 for Gradient Boosting Tree, 7.534 for ExtraTrees, 9.182 for CatBoosts, 8.489 for LightGBM, 5.703 for XGBoosts, and 5.759 for KNN. On MAPE, HLOA-CNN-BIGRU is 17.355, the lowest of all models, and is lower than 43.368 for Decision Tree, 26.692 for Random Forest, 27.83 for AdaBoosts, 27.185 for Gradient Boosting Tree, 27.082 for ExtraTrees, 27.276 for CatBoosts, 32.406 for LightGBM, 26.013 for XGBoosts, and 25.826 for KNN. In the R^2 index, HLOA-CNN-BIGRU reaches 0.952, which is higher than 0.656 for Decision Tree, 0.803 for Random Forest, 0.861 for AdaBoosts, 0.874 for Gradient Boost Tree, 0.84 for ExtraTrees, 0.784 for CatBoosts, 0.809 for LightGBM, 0.919 for XGBoosts, and 0.909 for KNN.

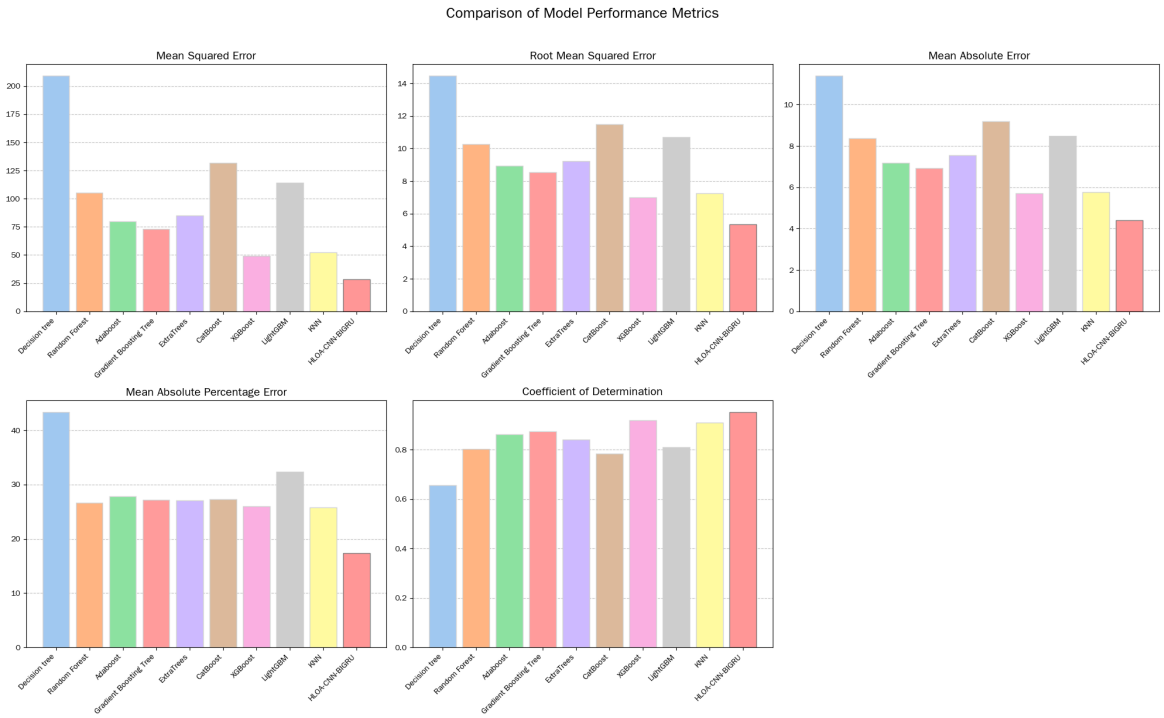


Figure 4. Tthe comparison bar chart of the indicators of each model

In order to solve the problems of query complexity, document noise, high cost, and lag in traditional manual evaluation of retrieval quality in retrieval-enhanced generative (RAG) systems, a retrieval quality prediction model integrating lizard optimization algorithm (HLOA), convolutional neural network (CNN), and bidirectional gated recursive unit (BIGRU) is proposed in this paper. Correlation analysis showed that there was a strong positive correlation between search level and search usefulness score, that is, the higher the search level, the better the usefulness of search results; There is a strong negative correlation between query complexity and retrieval usefulness score, and higher query complexity indicates reduced retrieval usefulness. Performance comparison experiments show that the HLOA-CNN-BIGRU model outperforms eight comparison models such as decision tree, random forest and Adaboost in key indicators. Not only are the MSE, RMSE, MAE, and MAPE error indices significantly reduced, but the determination coefficient R^2 reaches 0.952, proving the more accurate and stable prediction ability of retrieval quality.

DOI: 10.54254/2755-2721/2025.27800

This paper innovatively combines the parameter optimization advantages of HLOA with the feature extraction capabilities of CNN and BIGRU to provide a new model architecture for the retrieval quality prediction of RAG system; It effectively solves the pain point that traditional manual evaluation can't meet real-time optimization, and can provide a basis for dynamic adjustment of retrieval strategy in RAG system by accurately predicting the retrieval quality, thus improving the accuracy and reliability of content generated in scenarios such as intelligent Q&A and knowledge assistant, and promoting the more efficient implementation of RAG technology in practical applications.

5. Conclusion

In this paper, the improved least squares support vector machine (SBO-LSSVM) algorithm in the state optimization algorithm is applied to the parameter optimization and quality prediction of metal selective laser melting (SLM) process. The research firstly identifies the key influencing factors through the correlation analysis of indicators. From the correlation heat map, it can be observed that the correlation color between laser power, scanning speed, and layer thickness and the relative density of the part is darker. This directly proves the key adjustment role of these three parameters on the relative density of parts in SLM process, and provides a clear direction for subsequent parameter optimization. To verify the performance of SBO-LSSVM, experiments use decision tree, random forest, XGBoost, and LightGBM as control models. The results showed that the SBO-LSSVM performed best among all evaluated metrics: its MSE, RMSE, MAE and MAPE were 1.089, 1.043, 0.766 and 0.867, respectively, which were the smallest values among all models. Compared with decision trees (1.952, 1.397, 1.031, 1.159), random forests (2.195, 1.482, 1.002, 1.13), and LightGBM (2.017, 1.42, 1.018, 1.149), SBO-LSSVM significantly reduced the prediction error, especially MSE, which was about 24.6% lower than suboptimal XGBoost (1.445).

The 0.855 of SBO-LSSVM is also better than all control models in terms of the R^2 index of data fitting ability. The R^2 of Decision Tree, Random Forest, XGBoost, and LightGBM are 0.786, 0.751, 0.8, and 0.764, respectively, which fully demonstrates its stronger ability to capture actual data patterns of SLM processes. Although XGBoost performs relatively well on MAE (0.819) and R^2 (0.8), it is still not as good as SBO-LSSVM; Random forests exhibit weaknesses of highest MSE and lowest R^2, further confirming the combined advantages of SBO-LSSVM in terms of prediction accuracy and fit. This research not only provides more accurate quality prediction and parameter optimization tools for SLM technology, effectively reduces production losses and quality defects caused by unreasonable parameters, but also provides a feasible path for the optimization and upgrading of machine learning models in the field of additive manufacturing, which has important practical value for promoting SLM technology to high-precision industrial application.

References

[1] Huly, O., Carmel, D., & Kurland, O. (2025). Predicting RAG performance for text completion. *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[2] Zhao, S., et al. (2024). Towards understanding retrieval accuracy and prompt quality in RAG systems. *arXiv preprint arXiv:2411.19463*.

[3] Veturi, S., et al. (2024). RAG based question-answering for contextual response prediction system. *arXiv preprint arXiv:2409.03708*.

[4] Chan, C.-M., et al. (2024). RQ-RAG: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.

[5] Shi, Y., et al. (2024). Enhancing retrieval and managing retrieval: A four-module synergy for improved quality and efficiency in RAG systems. *arXiv preprint arXiv:2407.10670*.

DOI: 10.54254/2755-2721/2025.27800

[6] He, J., et al. (2025). Context-guided dynamic retrieval for improving generation quality in RAG models. *arXiv preprint arXiv:2504.19436*.
[7] Jiang, Z., Ma, X., & Chen, W. (2024). Longrag: Enhancing retrieval-augmented generation with long-context llms. *arXiv preprint arXiv:2406.15319*.
[8] Ampazis, N. (2024). Improving RAG quality for large language models with topic-enhanced reranking. In *IFIP International Conference on Artificial Intelligence Applications and Innovations* (pp. 10470–10490). Cham: Springer Nature Switzerland.
[9] Yang, X., et al. (2024). Crag-comprehensive rag benchmark. *Advances in Neural Information Processing Systems, 37*, 10470–10490.
[10] Zhang, Z., Fang, M., & Chen, L. (2024). Retrievalqa: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. *arXiv preprint arXiv:2402.16457*.