Proceedings of the 3rd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2025.29149

Optimization Prediction of RAG Retrieval Performance Based on Transformer-BiGRU Regression Algorithm

Qi He

Department of Industrial Engineering, Tsinghua University, Beijing, China
moziofmoon@163.com

Abstract. The rolling bearings of the train bogie system, as the core basic components of rail transit trains, are widely used in the bogie system and are key components to ensure the safety of train operation and the reliability of train operation. With the in-depth advancement of intelligent rail transit, high-speed trains are gradually developing towards high speed, high reliability and long endurance, and the safety requirements for the rolling bearings of the bogie system are becoming increasingly strict. This paper proposes a classification algorithm for rolling bearings in train bogie systems based on KELM-LSTM-Transformer. During the research process, data preprocessing and data analysis are carried out first, followed by correlation analysis, and then this algorithm is compared with multiple machine learning algorithms. From the comparison results, it can be seen that the KELM-LSTM-Transformer significantly outperforms other single models and integrated models in all key evaluation dimensions Its Accuracy is 0.855, Recall is 0.855, Precision is 0.851, and F1 value is 0.847, all of which are the highest among all models. Not only does it far exceed models with relatively weak performance such as the BP neural network with an F1 value of only 0.641 and the GBDT with an F1 value of 0.732, It also outperforms better-performing models such as CatBoost with an Accuracy of 0.811 and an F1 value of 0.792, as well as XGBoost with an Accuracy of 0.792 and an F1 value of 0.795. In terms of the AUC metric for measuring the model's discrimination ability, the algorithm's value of 0.983 is also ahead of the integrated models such as CatBoost with an AUC of 0.974 and XGBoost with an AUC of 0.97, demonstrating a stronger ability to distinguish positive and negative samples. The proposal of this algorithm provides an effective technical solution for improving the accuracy of condition monitoring and fault classification of rolling bearings in the bogies of rail transit trains, which is of great practical significance for ensuring the safe and stable operation of high-speed trains and promoting the intelligent upgrade of the rail transit field.

Keywords: Transformer, BiGRU, RAG retrieval

1. Introduction

The retrieval stage of the retrieval Enhancement generation (RAG) system aims to find information related to the user's query from a large amount of data. This information will be used to generate responses subsequently [1]. The scale of data is constantly expanding, and the types of data are becoming increasingly complex. The traditional keyword matching retrieval method cannot well

© 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: 10.54254/2755-2721/2025.29149

capture the deep semantic connection between queries and documents [2]. This leads to insufficient relevance of the retrieved information. This situation will affect the overall performance of the RAG system. Therefore, correlation prediction has become the key to improving the retrieval quality of the RAG system. Researchers need to build effective predictive models to accurately determine the degree of correlation between queries and documents. This is currently an important research direction in this field [3].

Machine learning algorithms have the ability to learn data and extract features. This ability makes it very important in the correlation prediction of the retrieval stage in the RAG system. Traditional machine learning algorithms include support vector machines, logistic regression, etc. These algorithms require manual design of features and then learning [4]. They can initially complete correlation prediction, but they do not perform well when dealing with complex semantic information. The emergence of deep learning algorithms has brought new approaches to solving this problem. Convolutional neural networks, recurrent neural networks, etc. all belong to deep learning algorithms. They can automatically extract deep semantic features from the data. This has significantly improved the accuracy of correlation prediction. However, these deep learning models still have deficiencies in capturing long-distance dependencies and parallel processing [5].

There are some problems with the existing models, which affect the effect of correlation prediction in the retrieval link of the RAG system. To improve these problems and enhance the prediction performance, this paper proposes the Transformer-BiGRU regression model. The Transformer model has a self-attention mechanism. This mechanism can help it capture long-distance dependencies in the text. It can also comprehensively mine the associated information between queries and documents. The BiGRU model has a very good ability to process sequence information. It can effectively extract temporal features from the text. The Transformer-BiGRU regression model combines the advantages of these two models. It can predict the correlation between queries and documents more accurately, which will help improve the performance of the RAG system's retrieval process.

2. Data sources

This dataset is a correlation prediction dataset for the retrieval stage of the retrieval Enhancement Generation (RAG) system, containing a total of 725 sample records. Each record corresponds to a matching instance of a query and a retrieved document fragment, and it includes four major types of feature variables and one core predictor variable. The query feature variables include query length, query complexity and whether they contain professional terms. Document feature variables include document fragment length, chapter importance, number of days ago, and source reliability. The matching feature variables include keyword overlap rate, semantic similarity, the number of entity matches, and whether there is an exact phrase match. The retrieval environment feature variables include retrieval ranking and methods. Methods are represented by 0, 1, and 2 respectively for BM25, vector, and mixed retrieval. The core predictor variable is the correlation score, with a numerical range from 0 to 100.

DOI: 10.54254/2755-2721/2025.29149

Table 1. The statistical results of the dataset

Query length	Query has special terms	Doc length	Doc section importance	Doc age days	Entity matching count	Has exact phrase match	Retrieval rank	Retrieval method	Relevance score
56	0	474	4	573	3	0	16	2	15.7
97	1	290	4	48	1	0	6	2	20.7
19	0	329	2	1150	4	1	6	2	45.1
76	0	237	3	222	7	0	3	2	26.2

Correlation analysis was conducted on each variable, and a correlation heat map was drawn, as shown in Figure 1. From the results of the correlation analysis, it can be known that the coefficient between the query length and the correlation score is -0.054097, which is close to 0, indicating that the correlation between the two is very weak. The coefficient of query complexity is 0.054704. A weak positive correlation indicates that when the query complexity is slightly higher, the correlation score may increase slightly. The semantic similarity coefficient is 0.417744. This relatively high positive value indicates a strong positive correlation. The more similar the semantics are, the higher the correlation score. The coefficient of the number of entity matches is 0.261668. A positive correlation indicates that an increase in the number of matched entities helps to enhance the correlation. The search ranking coefficient is -0.096307, showing a negative correlation, which means that the lower the ranking, the lower the correlation score. The coefficient of the number of days since the document is -0.081905 is also negative, indicating that the older the document, the worse its relevance. Overall, the direction and degree of influence of different variables on the correlation score vary.

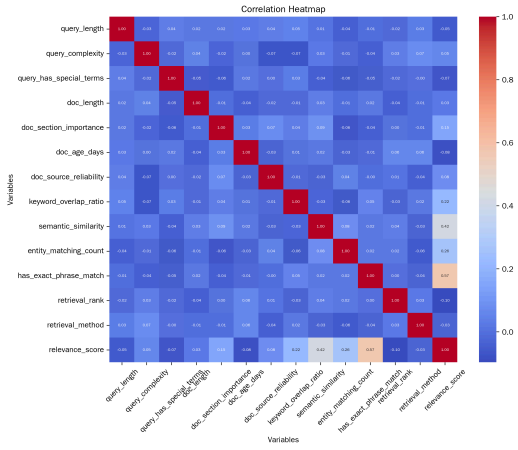


Figure 1. Correlation analysis

3. Method

3.1. Transformer

The Transformer algorithm is a deep learning model based on the self-attention mechanism. It breaks away from the sequence dependence of traditional recurrent neural networks and adopts parallel computing to enhance efficiency [6]. The core of the model includes an encoder and a decoder. The encoder is responsible for converting the input sequence into a vector representation

DOI: 10.54254/2755-2721/2025.29149

containing semantic information, and the decoder generates the target sequence based on this vector [7]. The self-attention mechanism is a key component that enables the model to calculate the degree of association between each element and all other elements in the sequence when processing it, thereby capturing global dependencies and enhancing the ability to utilize context information.

In terms of specific structure, both the encoder and the decoder are composed of multiple identical layers stacked together. Each layer contains multi-head attention mechanisms and feedforward neural networks. Multi-head attention captures the correlation information in the sequence from different dimensions through multiple parallel attention heads, enhancing the model's understanding of complex semantics. Feedforward neural networks independently process the vectors output by each attention mechanism, further extracting and transforming features. Residual join and layer normalization techniques are introduced into each layer, effectively alleviating the vanishing gradient problem during the training process, accelerating model convergence, and ensuring that the model can still stably learn under deep stacking [8]. The network structure of the Transformer is shown in Figure 2.

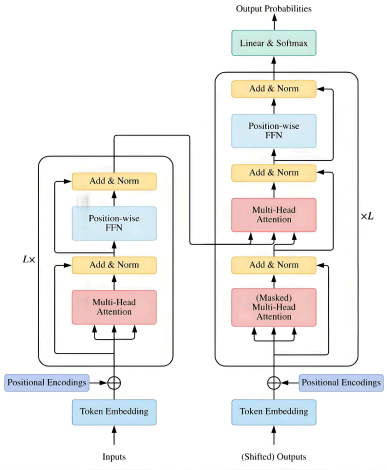


Figure 2. The network structure of the Transformer

3.2. BiGRU

The BiGRU algorithm, whose full name is Bidirectional Gated Recurrent Unit, is a deep learning model improved on the basis of GRU (Gated Recurrent Unit). It is mainly used to process sequential data. Its core advantage lies in its ability to simultaneously utilize the past and future information of the sequence, solving the limitation that traditional unidirectional recurrent neural networks can only rely on historical data [9]. The model consists of two GRU units in opposite directions. One processes the data in sequence from left to right to capture the features of past moments. Another one processes the data from right to left in sequence, captures the features of future moments, and finally combines the outputs from both directions to obtain a more comprehensive sequence semantic representation. The network structure of BiGRU is shown in Figure 3.

DOI: 10.54254/2755-2721/2025.29149

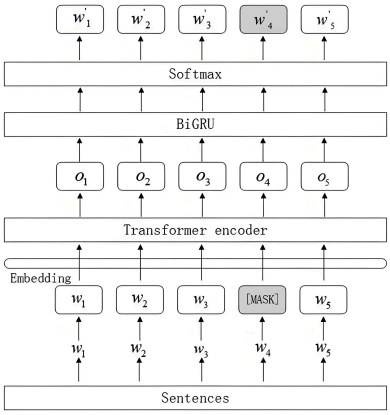


Figure 3. The network structure of BiGRU

BiGRU inherits the gating mechanism of GRU, including the reset gate and update gate. The reset gate is used to control whether historical information is ignored, helping the model focus on the input at the current moment. The update gate is used to determine how much historical information to retain to the current state, achieving long-term memory of important information. This gating structure can effectively alleviate the vanishing gradient problem in the training of recurrent neural networks, enabling the model to maintain good performance when processing long sequence data [10].

3.3. Transformer-BiGRU

The Transformer-BigRU algorithm is a deep learning model that combines the advantages of Transformer and BiGRU, mainly used for processing sequential data. It can capture the global dependencies of sequences by leveraging the self-attention mechanism of Transformer, and at the same time, utilize the bidirectional characteristics of BiGRU to capture local time series information, making up for the deficiencies of a single model when processing sequence data. Models typically first encode the input sequence through the Transformer module to obtain feature vectors containing global semantics, and then input these vectors into the BiGRU module for further processing.

In specific operation, the multi-head attention mechanism of the Transformer module mines the correlations among sequence elements from multiple dimensions to generate global features. The BiGRU module processes global features from left to right and from right to left respectively through two reverse GRU units, captures local temporal dependencies, and generates a more comprehensive sequence representation. This structure enables the model to not only grasp the overall semantics but also pay attention to local temporal details, thereby enhancing the processing effect.

4. Result

In terms of the parameter Settings of the model, the training set accounts for 70% of the dataset. The number of channels in the model sequence input layer is equal to the input feature dimension. The maximum position encoding of the position embedding layer is 512. The number of heads in the self-attention layer is 4, the number of channels for each head key is 32, and the total number of key channels is 128. It contains two self-attention layers, a forward GRU layer of dimension 6, a reverse GRU layer of dimension 10, as well as a concatenation layer, a ReLU activation layer, a Dropout

DOI: 10.54254/2755-2721/2025.29149

layer of 0.01 scale, an index layer and a fully connected layer. The Adam optimizer is used for training, with a maximum of 200 rounds and a batch size of 256. The dataset is shuffled in each round. The initial learning rate is 0.01, and it decayed by 0.1 times after 80 rounds. The L2 regularization coefficient is 0.001, and the gradient clipping threshold is 10.

In this paper, Decision tree, Random Forest, AdaBoost, GBDT, ExtraTrees and XGBoost are used as comparison models, and MSE, RMSE, MAE, MAPE and R^2 are used to evaluate the optimization effects of the models. The prediction results of each model are shown in Table 2. Output the bar chart comparison of each indicator as shown in the figure 4.

Table 2. The results of the comparative experiment

Model	MSE	RMSE	MAE	MAPE	R^2
Decision tree	48.051	6.932	5.46	22.159	0.372
Random Forest	35.804	5.984	4.878	18.579	0.565
AdaBoost	37.297	6.107	4.728	18.073	0.509
GBDT	34.903	5.908	4.801	19.43	0.541
ExtraTrees	33.04	5.748	4.552	17.329	0.551
XGBoost	36.644	6.053	4.856	20.243	0.49
Transformer-BiGRU(Our model)	31.322	5.597	4.43	18.207	0.598

From the perspective of various evaluation indicators, the overall performance of the Transformer-BiGRU model proposed in this paper is superior to that of the six models, namely Decision Tree, Random Forest, AdaBoost, GBDT, ExtraTrees and XGBoost. Among the metrics for measuring prediction errors, the mean square error (MSE) of Transformer-BiGRU is 31.322. This value is the lowest among all comparison models and has a significant advantage over models such as 48.051 of decision trees and 35.804 of random forests. This result indicates that the square error between the predicted value and the true value of Transformer-BiGRU is the smallest. The root mean square error (RMSE) corresponding to Transformer-BiGRU is 5.597, which is also lower than that of other comparison models. This result further indicates that the overall error level of Transformer-BiGRU is lower. In terms of the Mean Absolute Error (MAE) metric, the value of Transformer-BiGRU is 4.43 and this value is smaller than that of other comparison models. This means that the mean absolute error of Transformer-BiGRU is smaller and the absolute deviation between its predicted value and the true value is more controllable. On the Mean Absolute percentage error (MAPE) metric, the value of Transformer-BiGRU is 18.207, which is slightly higher than that of ExtraTrees at 17.329 and AdaBoost at 18.073, indicating a certain deficiency. However, 18.207 is still lower than the MAPE values of the four models: decision tree, Random Forest, GBDT and XGBoost. In terms of the coefficient of determination (R^2) index for measuring the goodness of fit of the model, the value of Transformer-BiGRU is 0.598 and this value is the highest among all the comparison models. This result indicates that Transformer-BiGRU has a stronger ability to interpret data variations and a better fitting effect.

DOI: 10.54254/2755-2721/2025.29149

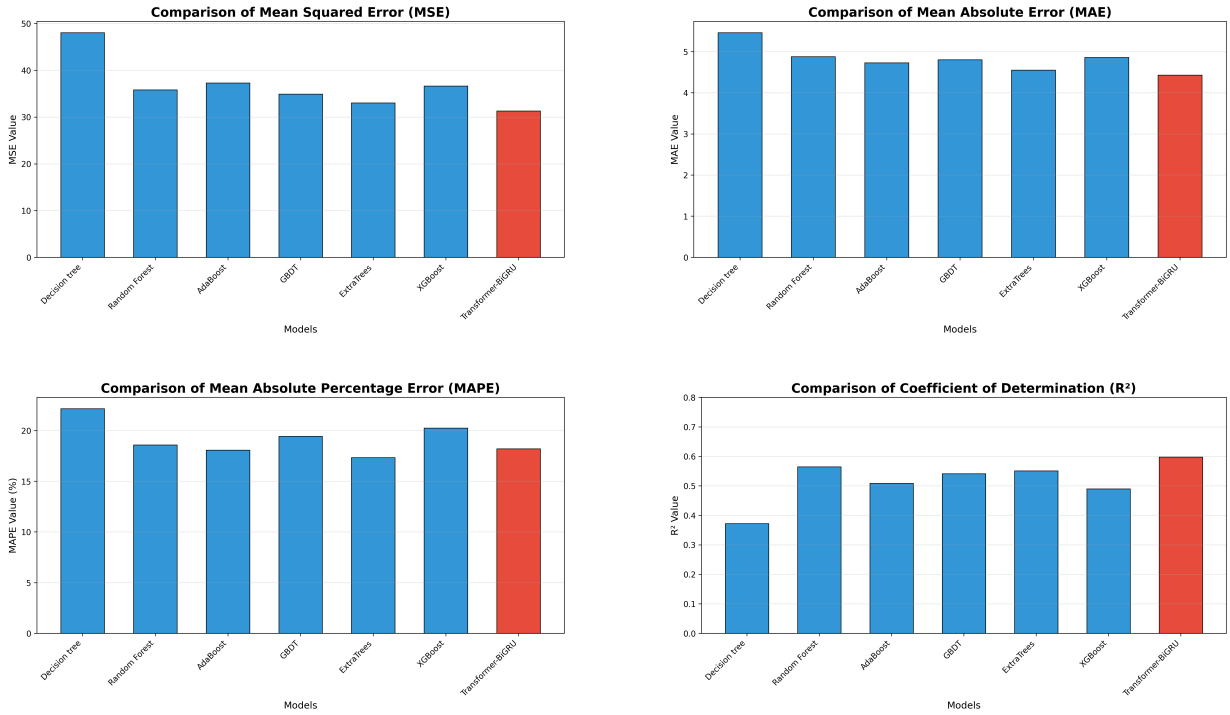


Figure 4. The bar chart comparison of each indicator

The line graph of the predicted values - actual values of the test set of Transformer-BiGRU(Our model) is shown in the figure 5.

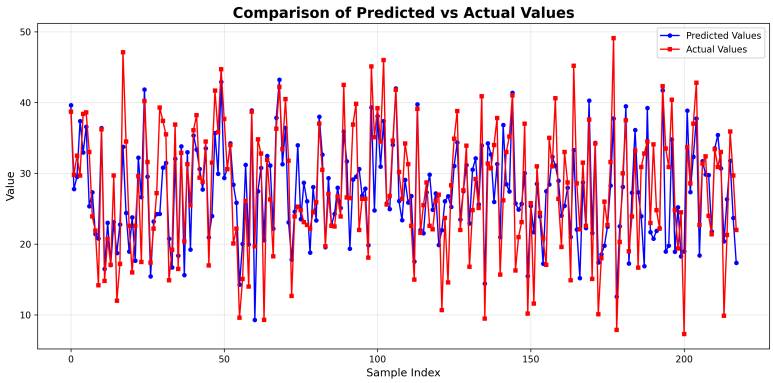


Figure 5. The line graph of the predicted values - actual values of the test set of Transformer-BiGRU

5. Conclusion

In the retrieval stage of the Retrieval Enhancement Generation (RAG) system, the traditional keyword matching method has deficiencies. This method is difficult to capture the deep semantic connection between the query and the document, ultimately resulting in insufficient relevance of the retrieved information. To solve this problem, this paper proposes the Transformer-BiGRU regression model, with the aim of optimizing the correlation prediction effect. Correlation analysis was carried out in the research, and comparative experiments of multiple machine learning algorithms were also conducted. It can be known from the experimental results that the Transformer-BiGRU model proposed in this paper performs better overall than the six models of decision tree, Random Forest,

DOI: 10.54254/2755-2721/2025.29149

AdaBoost, GBDT, ExtraTrees and XGBoost. In terms of the measurement index of prediction error, the mean square error (MSE) of the Transformer-BiGRU model is 31.322. This value is the lowest among all the comparison models and is significantly better than models such as 48.051 of decision tree and 35.804 of random forest. This result indicates that the square error between the predicted value and the true value of this model is the smallest. The root mean square error (RMSE) corresponding to the Transformer-BiGRU model is 5.597, which is also lower than that of other comparison models. This result further confirms that the overall error level of this model is lower. The mean absolute error (MAE) of the Transformer-BiGRU model is 4.43, which is smaller than that of the other comparison models. This means that the mean absolute error of this model is smaller, and the absolute deviation between its predicted value and the true value is easier to control. In terms of the Mean Absolute percentage error (MAPE) metric, the value of the Transformer-BiGRU model is 18.207, which is slightly higher than that of ExtraTrees at 17.329 and AdaBoost at 18.073. However, it is still lower than the MAPE values of the other four models, namely decision tree, Random Forest, GBDT and XGBoost. In terms of the coefficient of determination (R^2) index for measuring the goodness of fit of the model, the value of the Transformer-BiGRU model is 0.598, which is the highest among all the comparison models. This result indicates that this model has a stronger ability to interpret data changes and a better fitting effect. Overall, the Transformer-BiGRU regression model demonstrates superior performance in the correlation prediction of the RAG system retrieval stage in most key evaluation metrics. It is only slightly inferior to the ExtraTrees and AdaBoost models in the MAPE metric, and its overall prediction performance is more outstanding.

References

[1] Salemi, Alireza, and Hamed Zamani. "Evaluating retrieval quality in retrieval-augmented generation." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.
[2] Leng, Quinn, et al. "Long context rag performance of large language models." arXiv preprint arXiv: 2411.03538 (2024).
[3] Shi, Yunxiao, et al. "Enhancing retrieval and managing retrieval: A four-module synergy for improved quality and efficiency in rag systems." arXiv preprint arXiv: 2407.10670 (2024).
[4] Zhao, Shengming, et al. "Towards understanding retrieval accuracy and prompt quality in rag systems." arXiv preprint arXiv: 2411.19463 (2024).
[5] Jin, Bowen, et al. "Long-context llms meet rag: Overcoming challenges for long inputs in rag." arXiv preprint arXiv: 2410.05983 (2024).
[6] Mengmeng, Su, et al. "An Effective Retrieval Method to Improve RAG Performance." 2024 7th International Conference on Data Science and Information Technology (DSIT). IEEE, 2024.
[7] Tang, Yixuan, and Yi Yang. "Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries." arXiv preprint arXiv: 2401.15391 (2024).
[8] Jiang, Wenqi, et al. "Rago: Systematic performance optimization for retrieval-augmented generation serving." Proceedings of the 52nd Annual International Symposium on Computer Architecture. 2025.
[9] Oche, Agada Joseph, et al. "A systematic review of key retrieval-augmented generation (rag) systems: Progress, gaps, and future directions." arXiv preprint arXiv: 2507.18910 (2025).
[10] Luo, Kun, et al. "Does RAG Really Perform Bad For Long-Context Processing?." arXiv preprint arXiv: 2502.11444 (2025).