

# Health State Prediction of Traction Motor Bearings Based on BiLSTM-Transformer

Yumei Liu<sup>1, a</sup>, Zongyao Li<sup>1, b \*</sup>, Yanxuan Zhou<sup>1, c</sup>, Zhiyuan Zhang<sup>2, d</sup>,  
Ming Liu<sup>1, e</sup>

<sup>1</sup> College of Transportation, Jilin University, Changchun 130022, China;

<sup>2</sup> Geely Automobile Research Institute (Ningbo) Co. Ltd., Ningbo, CHINA.

<sup>a</sup>lymlls@163.com, <sup>b</sup>zongyao24@mails.jlu.edu.cn, <sup>c</sup>13843148137@163.com, <sup>d</sup>zy.z@geely.com,  
<sup>e</sup>2635458147@qq.com

**Abstract.** Bearings are core components of the traction system in high-speed trains, and ensuring their safe and reliable operation is essential for the normal functioning of the trains. Accurate health state prediction is critical to the safe operation and maintenance of traction motors in high-speed trains. For this reason, this paper first extracts time domain and frequency domain features from vibration signals, fuses degradation features using a Self-Organizing Map (SOM) neural network, and applies the Cumulative Sum (CUSUM) control chart method to construct a quantitative health indicator (HI). Then, Bidirectional Long Short-Term Memory (BiLSTM) network is employed to extract short-term features, while the Transformer is used to capture long-term dependencies, thereby building a health state prediction model aimed at improving prediction accuracy. Finally, the effectiveness of the proposed model is validated on the XJTU-SY bearing dataset.

**Keywords:** Traction motor bearings; Health indicator; BiLSTM; Transformer; Health state prediction.

## 1. Introduction

As a critical component of the traction motor in high-speed trains, bearings have a high failure rate and are often subject to progressive faults characterized by significant performance degradation. These failures are among the primary causes of traction motor malfunction and directly affect the overall performance and operational safety of high-speed trains [1]. Health state prediction, based on the evolving trend of degradation characteristics of traction motor bearings, enables timely maintenance and effectively reduces the occurrence of failures. However, due to the complex operating environment of traction motors, vibration signals from different components are prone to interference and overlap, making it difficult for a single feature to comprehensively reflect the health condition of the bearing. Moreover, these features are highly sensitive to changes in operating conditions. Therefore, accurate prediction of bearing health status is of great importance for ensuring the reliable operation of traction motors.

Health state prediction methods are mainly categorized into knowledge-driven, model-driven, and data-driven approaches. Knowledge-driven methods make full use of expert experience, but their effectiveness is limited due to difficulties in knowledge acquisition and a lack of real-time adaptability [2]. Model-driven approaches are suitable for simple mechanical systems with well-defined physical models, but it is challenging to build accurate models for components like traction motor bearings that operate under complex conditions [3]. With the development of artificial intelligence, data-driven methods have been widely adopted for health prediction [4]. In particular, deep learning-based approaches have rapidly advanced in the field due to their powerful nonlinear mapping and feature extraction capabilities [5]. Representative methods include Deep Belief Networks (DBN) [6], BiLSTM network [7], and Transformer [8]. BiLSTM can capture both past and future information in sequences and is especially effective for modeling short-term dependencies in time-series data. However, its ability to learn long-term dependencies is relatively limited. The Transformer, with its self-attention mechanism, can directly capture global

dependencies across time steps in the input sequence, making it well-suited for modeling long-range dependencies. Nevertheless, it is less effective at capturing short-term patterns, particularly when the amount of training data is limited.

To combine the advantages of BiLSTM and Transformer, this paper proposes a health state prediction model based on the BiLSTM-Transformer architecture. First, degradation features are extracted from bearing vibration signals. These features are then fused using a SOM neural network. A HI capable of quantitatively representing the performance degradation of bearings is CUSUM control chart method. By integrating the strengths of BiLSTM in extracting short-term features and Transformer in capturing long-term dependencies, a BiLSTM-Transformer-based HI prediction model is developed. The effectiveness of the proposed model is validated using the XJTU-SY bearing dataset.

### 1.1 Degradation Feature Extraction

This paper extracts time domain and frequency domain features from vibration signals to provide a more comprehensive representation of bearing operating conditions and improve the accuracy of health state prediction.

Let the vibration signal be denoted as  $x=[x_1, x_2, \dots, x_N]$ , and the sampling length as  $N$ . The formulas for calculating time-domain features are shown in Table 1.

Table 1. Time domain features

| Expression                                                     | Expression                                                  | Expression                                            | Expression                      |
|----------------------------------------------------------------|-------------------------------------------------------------|-------------------------------------------------------|---------------------------------|
| $F_1 = \max(x_i)$                                              | $F_5 = \frac{1}{N} \sum_{i=1}^N x_i^2$                      | $F_9 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{X})^3$    | $F_{13} = \frac{F_1}{F_3}$      |
| $F_2 = \frac{1}{N} \sum_{i=1}^N x_i$                           | $F_6 = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2}$               | $F_{10} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{X})^4$ | $F_{14} = \frac{F_1}{F_4}$      |
| $F_3 = \frac{1}{N} \sum_{i=1}^N  x_i $                         | $F_7 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{X})^2$        | $F_{11} = \frac{F_6}{F_3}$                            | $F_{15} = \frac{F_{10}}{F_8^4}$ |
| $F_4 = \left( \frac{1}{N} \sum_{i=1}^N \sqrt{ x_i } \right)^2$ | $F_8 = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{X})^2}$ | $F_{12} = \frac{F_1}{F_6}$                            | $F_{16} = \frac{F_9}{F_8^4}$    |

The expressions for calculating frequency domain features are shown in Table 2.

Table 2. frequency domain features

| Expression                                                | Expression                                                        | Expression                                                               | Expression                                                         |
|-----------------------------------------------------------|-------------------------------------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------|
| $S_1 = \frac{1}{K} \sum_{k=1}^K s(k)$                     | $S_4 = \frac{\sum_{k=1}^K (f_k - S_2)^2 s(k)}{\sum_{k=1}^K s(k)}$ | $S_7 = \frac{\sum_{k=1}^K [s(k) - \bar{S}]^3}{\sqrt{S_6}^3}$             | $S_{10} = \frac{\sum_{k=1}^K (f_k - S_2)^3 s(k)}{AC^3 K}$          |
| $S_2 = \frac{\sum_{k=1}^K f_k s(k)}{\sum_{k=1}^K s(k)}$   | $S_5 = \sqrt{\frac{\sum_{k=1}^K f_k^2 s(k)}{\sum_{k=1}^K s(k)}}$  | $S_8 = \frac{\sum_{k=1}^K [s(k) - \bar{S}]^4}{\sqrt{S_6}^3}$             | $S_{11} = \frac{\sum_{k=1}^K (f_k - S_2)^4 s(k)}{AC^4 K}$          |
| $S_3 = \frac{\sum_{k=1}^K f_k^2 s(k)}{\sum_{k=1}^K s(k)}$ | $S_6 = \frac{1}{K-1} \sum_{k=1}^K [s(k) - \bar{S}]^2$             | $S_9 = \sqrt{\frac{\sum_{k=1}^K (f_k - S_2)^2 s(k)}{\sum_{k=1}^K s(k)}}$ | $S_{12} = \frac{\sum_{k=1}^K \sqrt{ f_s - S_2 } s(k)}{\sqrt{ACK}}$ |

### 1.2 Degradation Feature Selection

The evaluation indicators for degradation feature selection include relevance, monotonicity, and robustness.

#### (1) Correlation

Correlation in the degradation process refers to the strength of the relationship between a feature and the degradation level of the equipment. The calculation is shown in equation (1).

$$Corr(X) = \frac{\left| n \sum_i x_i t_i - \sum_i x_i \sum_i t_i \right|}{\sqrt{\left[ n \sum_i x_i^2 - \left( \sum_i x_i \right)^2 \right] \left[ n \sum_i t_i^2 - \left( \sum_i t_i \right)^2 \right]}} \quad (1)$$

In the above equation,  $X = (x_1, x_2, \dots, x_n)$  denotes the time series of the degradation feature;  $n$  is the number of samples; and  $T = (t_1, t_2, \dots, t_n)$  represents the sampling time series.

### (2) Monotonicity

Monotonicity describes whether the relationship between a feature and the degradation process exhibits a consistent directional change. The calculation is shown in equation (2).

$$Mon(X) = \frac{\left| \sum_i \varepsilon(x_i - x_{i-1}) - \sum_i \varepsilon(x_{i-1} - x_i) \right|}{n-1} \quad (2)$$

Where  $\varepsilon(x) = \begin{cases} 1(x \geq 0) \\ 0(x < 0) \end{cases}$  is the unit step function.

### (3) Robustness

Robustness refers to the stability of a feature when subjected to data noise or environmental variations. The calculation is shown in equation (3).

$$Rob(X) = \frac{1}{n} \sum \exp \left( - \left| \frac{x_i - \tilde{x}}{x_i} \right| \right) \quad (3)$$

Where  $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)$  is the trend component of the feature time series.

By comprehensively considering the evaluation metrics of correlation, monotonicity, and robustness, a linear combination equation is established to construct a comprehensive evaluation index. This index is used to evaluate the features and select those that are sensitive to degradation, as shown in equation (4).

$$W = \omega_1 Corr(X) + \omega_2 Mon(X) + \omega_3 Rob(X) \\ s.t. \begin{cases} \omega_i \geq 0 \\ \sum_i \omega_i = 1 \end{cases} \quad (4)$$

Where  $W$  is the comprehensive evaluation index, and  $\omega_i (i=1,2,3)$  represents the weights of the evaluation metrics. In this paper, the entropy weight method is used to assign weights to the evaluation metrics [9]. The top 15 features with the highest comprehensive evaluation index  $W$  are selected as degradation-sensitive features.

## 1.3 Degradation Feature Fusion

The SOM is an unsupervised learning algorithm that can effectively learn the feature space of both normal and degraded states. Therefore, in this paper, feature fusion based on SOM is employed to generate a HI that reflects the degree of degradation. In addition, the CUSUM method is used to analyze the cumulative deviation in the time series, enabling more sensitive detection of anomalies in the data. The parameters of the SOM network are configured as follows: the input layer consists of 15 neurons, and a rectangular grid topology is adopted with a grid size of 10×10. The initial learning rate is set to 0.8, with an initial neighborhood radius of 5. The learning rate and neighborhood radius decay over 200 iterations, and the maximum number of training iterations is 2000.

To intuitively illustrate the evolution of the health state, the HI is divided into five ranges:  $[0, 0.1)$ ,  $[0.1, 0.25)$ ,  $[0.25, 0.45)$ ,  $[0.45, 0.7)$ , and  $[0.7, 1]$ , corresponding to the categories Healthy, Sub-healthy, Qualified, Warning, and Failure, respectively.

## 2. Health State Prediction Based on BiLSTM-Transformer

### 2.1 BiLSTM Model

The BiLSTM network consists of two independent LSTM layers. By simultaneously processing forward and backward sequences, it integrates past and future information, thereby enhancing the model's ability to capture bidirectional dependencies. The structures of the LSTM and BiLSTM networks are shown in Fig. 1(a) and Fig. 1(b), respectively. Compared with the traditional LSTM, the BiLSTM is more effective in capturing potential nonlinear relationships within the sequence.

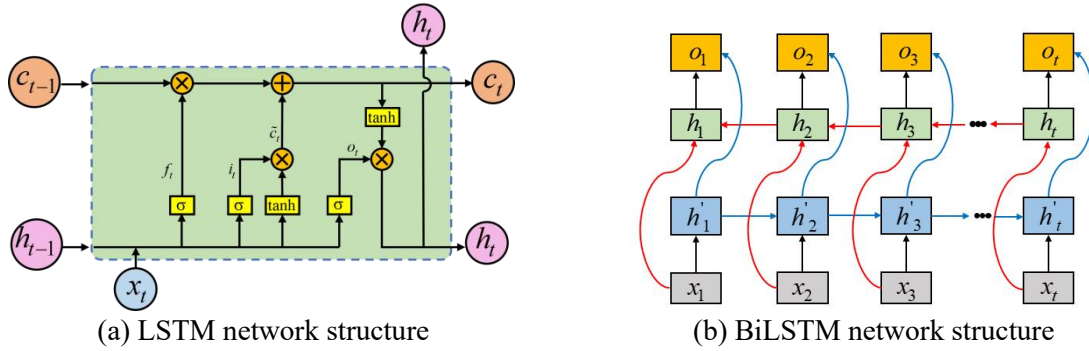


Fig. 1 LSTM and BiLSTM networks structure

### 2.2 Transformer Model

The Transformer is designed for sequence-to-sequence tasks, with its core concept being the complete abandonment of traditional recurrent structures in favor of an architecture built entirely on the Attention Mechanism. Compared with conventional neural networks, which suffer from low information transmission efficiency and limited parallelization capability on long sequences, the Transformer leverages the self-attention mechanism to capture dependencies between all positions in the input sequence without being constrained by sequence length, thereby significantly improving the efficiency of information transmission. The structure of the Transformer model is shown in Fig. 2.

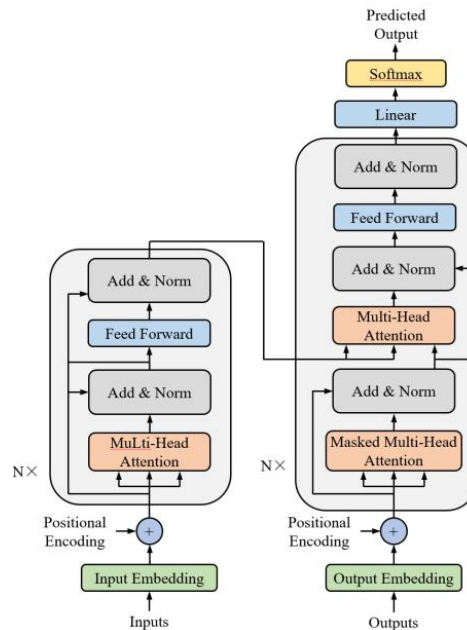


Fig. 2 Transformer model structure

### 2.3 BiLSTM-Transformer Health State Prediction Model

The HI prediction problem investigated in this paper falls under time series forecasting, which does not require step-by-step sequence generation or handling the historical information of the target sequence. Therefore, based on the classical Transformer model, the encoder and positional encoding components are retained, while the decoder is replaced with a linear fully connected layer to produce the prediction output. The structure of the proposed BiLSTM-Transformer model is shown in Fig.3. Its hyperparameters are optimized through network search tuning, as shown in Table 3.

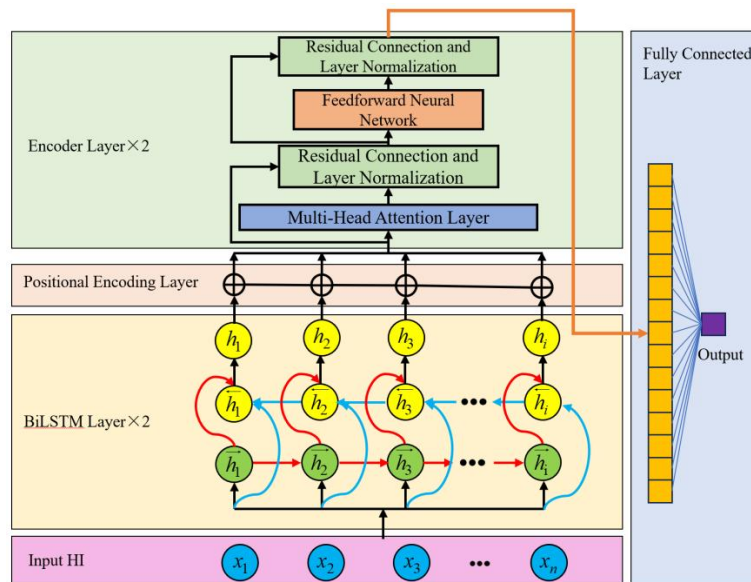


Fig. 3 Structure of the BiLSTM-Transformer prediction model

Table 3. Hyperparameters of the BiLSTM-Transformer model

| Model             | Parameters                              | Value                     |
|-------------------|-----------------------------------------|---------------------------|
| BiLSTM Layer      | Input Feature Dimension                 | 1                         |
|                   | Input Sequence Length                   | 32                        |
|                   | Number of Hidden Units (Unidirectional) | 64                        |
|                   | Output Feature Dimension                | 128                       |
|                   | Number of BiLSTM Layers                 | 2                         |
|                   | Dropout Rate                            | 0.2                       |
| Transformer Layer | Number of Encoder Layers                | 2                         |
|                   | Input Feature Dimension                 | 128                       |
|                   | Number of Attention Heads               | 8                         |
|                   | Positional Encoding                     | Sine and Cosine Functions |
|                   | Feature Dimension per Head              | 16                        |
|                   | Input Dimension of Feedforward Network  | 512                       |
|                   | Output Dimension of Feedforward Network | 128                       |
|                   | Activation Function                     | ReLu                      |
| Output Layer      | Dropout Rate                            | 0.2                       |
|                   | Input Feature Dimension                 | 128                       |
|                   | Output Feature Dimension                | 1                         |
| Overall           | Sequence Length                         | 32                        |
|                   | Sliding Step Size                       | 1                         |
|                   | Batch Size                              | 64                        |
|                   | Initial Learning Rate                   | 0.001                     |
|                   | Loss Function                           | Mean Squared Error (MSE)  |
|                   | Optimizer                               | Adam                      |

## 2.4 Model Evaluation Metrics

To evaluate the prediction performance of the model, the Mean Squared Error, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination ( $R^2$ ) are selected as evaluation metrics, as shown in Equations (5)–(8).

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (5)$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (6)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (7)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (8)$$

Where  $n$  is the number of prediction points,  $y_i$  represents the true values of the health indicator,  $\hat{y}_i$  denotes the predicted values of the health indicator, and  $\bar{y}$  indicates the mean of the health indicator values involved in the prediction. The values of MSE, RMSE, and MAE all fall within range  $[0, +\infty)$ , where smaller values indicate better prediction performance of the model. The  $R^2$  reflects the correlation between predicted and true values, with a theoretical range of  $[0, 1]$ . The closer the  $R^2$  value is to 1, the better the model's prediction performance. In this study,  $R^2$  is used to describe the prediction accuracy of the model.

## 3. Experimental Validation

The experimental data are obtained from the XJTU-SY rolling bearing accelerated life test dataset provided by Xi'an Jiaotong University [10]. The sampling duration for each instance is set to 1.28 seconds, with a sampling interval of 1 minute and a sampling frequency of 25.6 kHz, resulting in 32,768 data points per sample.

### 3.1 Health Indicator Prediction

The selection of the train and test sets is shown in Table 4. The prediction results and evaluation metrics are presented in Fig. 4 and Table 5, respectively.

Table 4. The selection of the train and test sets

| Operating Condition | Train dataset | Test dataset            |
|---------------------|---------------|-------------------------|
| Condition 1         | 1-1, 1-2      | 1-3, 1-4, 1-5           |
| Condition 2         | Not Trained   | 2-1, 2-2, 2-3, 2-4, 2-5 |
| Condition 3         | 3-1, 3-2      | 3-3, 3-4, 3-5           |

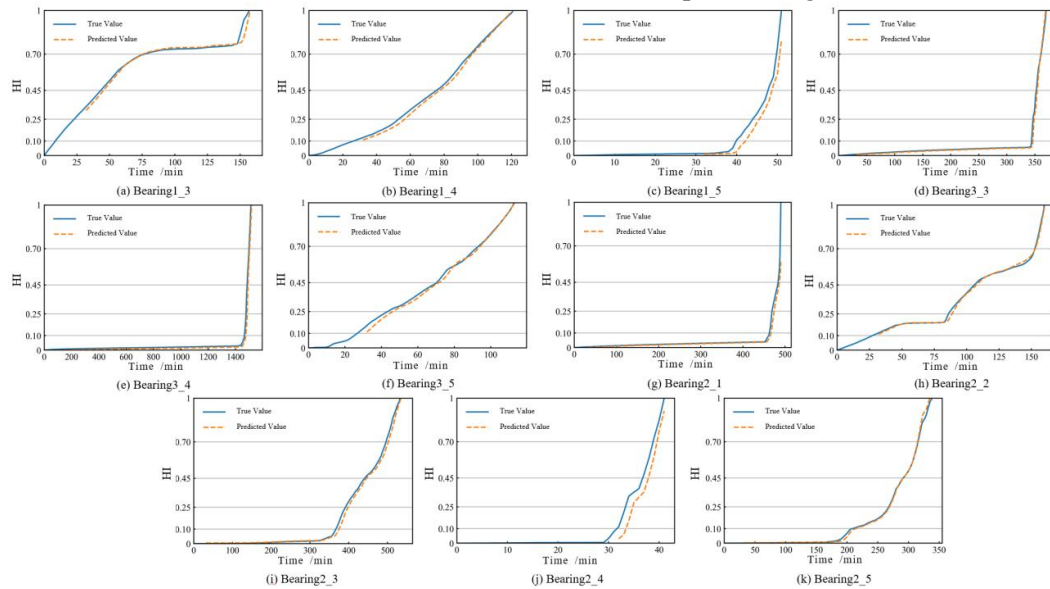


Fig. 4 Prediction results of the BiLSTM-Transformer model on the test set

Table 5. Evaluation metrics of the BiLSTM-Transformer model on the test set

| Operating Condition                           | Test set   | MSE     | RMSE    | MAE     | R <sup>2</sup> (%) |
|-----------------------------------------------|------------|---------|---------|---------|--------------------|
| Operating Condition 1                         | Bearing1_3 | 0.00015 | 0.01225 | 0.00966 | 99.175             |
|                                               | Bearing1_4 | 0.00016 | 0.01273 | 0.01003 | 99.381             |
|                                               | Bearing1_5 | 0.00646 | 0.08039 | 0.05998 | 89.874             |
| Operating Condition 3                         | Bearing3_3 | 0.00006 | 0.00775 | 0.00700 | 99.630             |
|                                               | Bearing3_4 | 0.00005 | 0.00707 | 0.00682 | 99.848             |
|                                               | Bearing3_5 | 0.00012 | 0.01095 | 0.00930 | 98.781             |
| Operating Condition 2<br>(Unseen in Training) | Bearing2_1 | 0.00018 | 0.01342 | 0.00849 | 92.437             |
|                                               | Bearing2_2 | 0.00020 | 0.01420 | 0.00860 | 99.250             |
|                                               | Bearing2_3 | 0.00012 | 0.01095 | 0.00922 | 98.551             |
|                                               | Bearing2_4 | 0.01279 | 0.11312 | 0.10742 | 88.521             |
|                                               | Bearing2_5 | 0.00010 | 0.01000 | 0.00801 | 99.309             |

As observed from Fig. 4 and Table 5, the model demonstrates excellent prediction performance on the three test bearings under operating conditions 1 and 3. Additionally, the model also achieves a high level of accuracy in predicting the bearings under condition 2, indicating that it maintains strong predictive capability even when applied to data from previously unseen operating conditions. According to the calculation, the average R<sup>2</sup> value on the test set is 96.796%, indicating that the proposed health state prediction model achieves an overall prediction accuracy of 96.796%, demonstrating excellent predictive performance.

### 3.2 Model Comparison Experiments

To verify the superiority of the proposed model, several comparison models are constructed, including the Feedforward Neural Network (FNN) [11], LSTM, BiLSTM, and Transformer model. All models are trained using the training set described in the previous section, and the HI data of Bearing 2-2 is used as the test set. The prediction results and evaluation metrics are shown in Table 6.

Table 6. Comparison of model prediction results

| Model              | MSE     | RMSE    | MAE     | R <sup>2</sup> (%) |
|--------------------|---------|---------|---------|--------------------|
| FNN                | 0.00266 | 0.05160 | 0.04387 | 87.890             |
| LSTM               | 0.00086 | 0.02931 | 0.02227 | 94.829             |
| BiLSTM             | 0.00078 | 0.02793 | 0.02104 | 96.625             |
| Transformer        | 0.00046 | 0.02152 | 0.01439 | 98.020             |
| BiLSTM-Transformer | 0.00020 | 0.01420 | 0.00860 | 99.250             |

The results show that the proposed BiLSTM-Transformer model outperforms the other comparison models across all evaluation metrics, fully validating its excellent performance in health state prediction tasks.

## 4. Summary

In this paper, the HI for traction motor bearings was constructed, and a health state prediction model based on the BiLSTM-Transformer architecture was proposed. Experimental validation on the XJTU-SY bearing dataset demonstrated that the model exhibits excellent prediction performance and generalization capability. Comparative analysis further confirmed that the proposed method achieves higher prediction accuracy. In the future, it can be integrated with model-based approaches to enhance the interpretability of the prediction results.

## Acknowledgements

This work was supported by the Natural Science Foundation of Jilin Province under Grant 20220101199JC, the Changchun Rail Transit Equipment Industry Major Science and Technology Special Project under Grant 2024GD04, and the Key Research and Development Projects of Jilin Science and Technology Department Foundation under Grant 20240302061GX.

## References

- [1] Liu Y, Li X, Zhang X, et al. Imbalanced deep transfer network for fault diagnosis of high-speed train traction motor bearings. *Knowledge-Based Systems*, 2024, 293: 111682.
- [2] Yin X, Rong Y, Li L, et al. Health state prediction method based on multi-featured parameter information fusion. *Applied Sciences*, 2024, 14(15): 6809.
- [3] Che C, Wang H, Fu Q, et al. Combining multiple deep learning algorithms for prognostic and health management of aircraft. *Aerospace Science and Technology*, 2019, 94: 105423.
- [4] Guo J, Wang Z, Li H, et al. A hybrid prognosis scheme for rolling bearings based on a novel health indicator and nonlinear Wiener process. *Reliability Engineering & System Safety*, 2024, 245: 110014.
- [5] Ye Z, Yu J. Health condition monitoring of machines based on long short-term memory convolutional autoencoder. *Applied Soft Computing*, 2021, 107: 107379.
- [6] Hu C H, Pei H, Si X S, et al. A prognostic model based on DBN and diffusion process for degrading bearing. *IEEE Transactions on Industrial Electronics*, 2019, 67(10): 8767-8777.
- [7] Guo J, Wang J, Wang Z, et al. A CNN-BiLSTM-Bootstrap integrated method for remaining useful life prediction of rolling bearings. *Quality and Reliability Engineering International*, 2023, 39(5): 1796-1813.
- [8] Kim S, Seo Y H, Park J. Transformer-based novel framework for remaining useful life prediction of lubricant in operational rolling bearings. *Reliability Engineering & System Safety*, 2024, 251: 110377.
- [9] Qiu G, Ye B, Gu Y, et al. A novel bearing remaining useful life prediction methodology with slope-based change point detection and WOA-Attention-BiLSTM model. *IEEE Sensors Journal*, 2025, 25(6), 10417-10431.
- [10] Wang B, Lei Y, Li N, et al. A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. *IEEE Transactions on Reliability*, 2018, 69(1): 401-412.
- [11] Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 1958, 65(6): 386-408.