Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 14, No. 2, 2025

A Bearing Fault Diagnosis Method Based on Fusion of CNN-BiLSTM-Transformer and Cross-Attention

Xun Yuan[1, a], Wei Wu[1]

[1]School of Mechanical Engineering, Xi'an Shiyou University, China
[a]yuanxun1221@163.com

Abstract: To address the limitations of traditional diagnostic models in achieving high accuracy in rolling bearing fault diagnosis and their inadequate extraction of vibration signal features across both frequency and time domains, this paper introduces a fault diagnosis approach that integrates multiple neural networks with a cross-attention mechanism. Initially, one-dimensional fault signals undergo processing through the Fast Fourier Transform (FFT) and Variational Mode Decomposition (VMD) techniques to obtain spectral information and time-domain features, respectively. The combined application of FFT and VMD facilitates the efficient extraction of multi-scale features from the signals. Subsequently, a CNN-Transformer network is utilized to capture the spatial characteristics of the pre-processed multi-scale fault signal features. This is followed by the deployment of a Bidirectional Long Short-Term Memory (BiLSTM) network to extract sequential information, with a particular emphasis on pivotal temporal features. Additionally, the integration of the Transformer network further enhances feature extraction and fusion capabilities. Ultimately, a cross-attention mechanism is implemented to seamlessly integrate time-domain and frequency-domain features, thereby bolstering the model's performance and generalization capacity for fault classification tasks. Experimental findings validate the proposed model's diagnostic effectiveness, achieving an accuracy and recall rate exceeding 99% in rolling bearing fault diagnosis.

Keywords: Rolling Bearing; Convolutional Neural Network (CNN); Bidirectional Long Short-Term Memory (BiLSTM); Cross-Attention; Transformer.

1. Introduction

In the operation of industrial equipment, rolling bearings, as a core component, play a vital role in ensuring the safe and reliable operation of machinery. The detection of bearing faults is therefore of paramount importance. Traditional fault diagnosis methods rely heavily on expert experience and basic signal processing techniques, which often struggle to effectively extract and analyze the complex characteristics of vibration signals, resulting in relatively low diagnostic accuracy. With the rapid development of deep learning technology, fault diagnosis methods based on deep neural networks have gradually demonstrated their superiority. However, how to fully utilize both the frequency-domain and time-domain features of vibration signals remains a significant research challenge.

In the field of vibration signal extraction and analysis, Fast Fourier Transform (FFT) and Variational Mode Decomposition (VMD) have been widely adopted as the primary methods. For instance, Reference [3] introduces a fault diagnosis technique that combines parameter-optimized Variational Mode Decomposition (VMD) with Convolutional Neural Networks (CNN). Reference [4] proposes a method based on VMD optimized by the Majorizing Sparrow Search Algorithm (MSSA-VMD). Reference [5] presents a rolling bearing fault diagnosis approach using Improved Multiscale Sample Entropy (IMSE) and parameter-optimized VMD. Reference [6] employs VMD to extract features from main bearing vibration data and then applies deep neural network-based fault diagnosis methods. Reference [7] demonstrates a technique for gearbox fault identification using Adaptive Variational Mode Decomposition. Reference [8] introduces a method for early fault detection in wind turbine gearboxes using Zoom-FFT-CEEMD combined with wavelet packet denoising. Reference [9] applies Hilbert Transform combined with cepstral analysis to process vibration data from gearboxes, aiming to identify and extract fault feature frequencies for effective diagnosis of wind turbine gearbox faults. In contrast, Reference [10] proposes an innovative method that combines the visualization of VMD signals with deep learning neural networks for precise bearing fault diagnosis.

With the continuous development of deep learning, various deep learning models have been extensively applied in bearing fault diagnosis. Specifically, Reference [11] describes a rolling bearing fault recognition strategy that integrates Ensemble Empirical Mode Decomposition (EEMD) with Convolutional Neural Network-Support Vector Machine (CNN-SVM). Reference [12] uses Backpropagation (BP) neural networks for bearing fault identification. Reference [13] introduces a fault diagnosis method based on Empirical Mode Decomposition (EMD) and optimized BP neural networks. Reference [14] proposes a deep neural network model that combines Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory networks (BiLSTM), and Attention Mechanism (AM) for early fault diagnosis. Reference [15] describes a fault diagnosis technique using R-vine Copula models combined with Dynamic Bayesian Networks. Reference [16] presents a gearbox fault diagnosis scheme based on multi-sensor information fusion and dual-stream Convolutional Neural Networks (CNN).

To enhance the accuracy of bearing fault diagnosis, this paper proposes a method that combines Fast Fourier Transform (FFT) and Variational Mode Decomposition (VMD) for data preprocessing, followed by a deep learning model integrating CNN, BiLSTM, Transformer, and Cross-Attention mechanisms (CNN-BiLSTM-Transformer-CrossAttention). Initially, one-dimensional fault signals are

preprocessed using FFT and VMD to obtain spectral information and multi-scale time-domain features. FFT effectively captures the frequency-domain characteristics of signals, while VMD aids in extracting multi-scale time-domain features. The combination of FFT and VMD allows for the comprehensive extraction of useful information from the signals.

In the feature extraction stage, Convolutional Neural Networks (CNN) and Transformer networks are employed to perform convolution and pooling operations on the preprocessed features to extract spatial features. Meanwhile, Bidirectional Long Short-Term Memory networks (BiLSTM) are used to learn sequence information and capture important temporal features. The introduction of Transformer networks enhances the features extracted by BiLSTM. Finally, Cross-Attention mechanisms are applied to fuse the features, further improving the extraction and integration capabilities of the model. This process enables the effective fusion of time-domain and frequency-domain features, thereby enhancing the diagnostic performance and generalization ability of the model.

2. Design of the CNN-BiLSTM-Transformer Model

Before feeding one-dimensional fault signals into the model, they must undergo preprocessing steps. These steps typically involve decomposing the signals in both the time and frequency domains to extract relevant feature information. In the frequency domain, although Fast Fourier Transform (FFT) can rapidly extract the spectral features of vibration signals and identify the main frequency components, it only provides global spectral information and is unable to capture the local time-varying characteristics of the signals.

Variational Mode Decomposition (VMD), as an adaptive signal decomposition technique, can effectively decompose complex vibration signals into multiple modal components with clear physical meanings. Each modal component reflects the signal's characteristics at different time scales. By combining FFT and VMD, it is possible to obtain both global spectral information and multi-scale local time-domain features, thereby achieving comprehensive feature extraction from the signals.

This multi-scale feature extraction method can more accurately capture the complex features of rolling bearing fault signals, thereby enhancing the accuracy and robustness of fault diagnosis. Moreover, the integration of frequency-domain and time-domain features enables better identification of different types and severity levels of faults, meeting the diagnostic requirements of complex industrial environments.

Based on the initial signal processing, the decomposed results are superimposed and processed using two network structures: CNN-Transformer and BiLSTM-Transformer, to extract features in both the spatial and temporal domains. Subsequently, Cross-Attention mechanisms are utilized to integrate the time-domain and frequency-domain features. In this manner, the model can compute and assign appropriate attention weights to different features, allowing it to focus more on key features and thereby enhancing its overall performance and generalization ability. The model structure is shown in Figure 1.

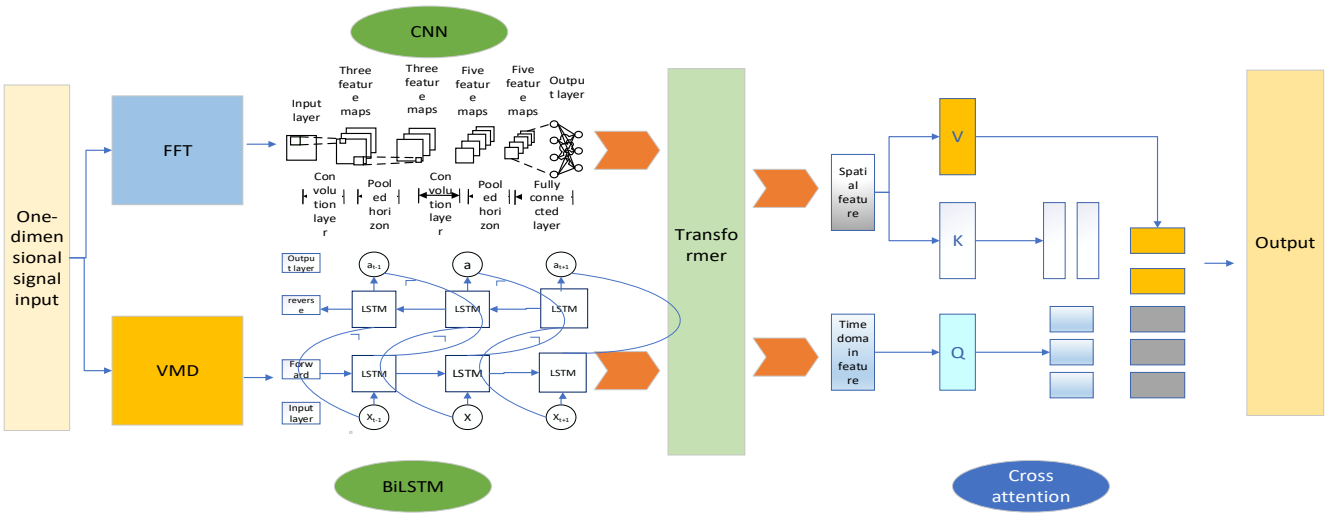

Figure 1. Model Architecture Diagram

2.1. Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) are a widely used deep learning architecture in the field of signal processing. The structure of the model is shown in Figure 2. The core concept of CNNs is to extract local features from input data through convolution operations and to progressively abstract and integrate these features via multiple convolutional layers, thereby obtaining high-level representations of the input data. In the context of rolling bearing fault diagnosis, the spectral information and multi-scale time-domain features of vibration signals can be regarded as two-dimensional or one-dimensional signal images. By sliding convolutional kernels over the input data, CNNs can efficiently capture and extract spatial feature information from the signals.

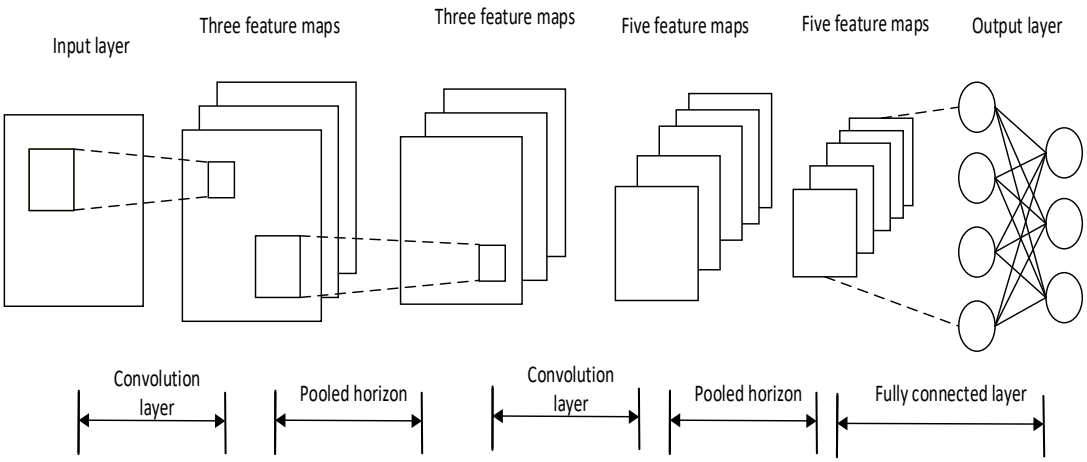

Figure 2. Convolutional Neural Network Model Diagram

Input Layer: This layer accepts preprocessed vibration signal features, such as feature maps obtained through Fast Fourier Transform (FFT) and Variational Mode Decomposition (VMD).

Convolutional Layer: Multiple convolutional kernels are applied to the input features to perform convolution operations. This process extracts different local features from the input data, capturing the spatial characteristics of the signals.

Pooling Layer: This layer reduces the dimensionality and computational complexity of the data by employing techniques such as max pooling or average pooling while retaining key features. The pooling operation helps to maintain the essential information while reducing the spatial size of the feature maps.

Fully Connected Layer: This layer transforms the high-level features extracted by the convolutional and pooling layers into a classification space. It plays a crucial role in mapping the extracted features to the output classes, enabling the final fault classification task.

Output Layer: The output layer utilizes the Softmax function to output the probabilities of different fault categories, thereby realizing fault classification. The Softmax function normalizes the output values into a probability distribution, making it suitable for multi-class classification tasks.

This architecture effectively extracts and integrates the spatial features of vibration signals, thereby enhancing the accuracy of fault diagnosis.

2.2. Bidirectional Long Short-Term Memory Network (BiLSTM)

The Bidirectional Long Short-Term Memory network (BiLSTM) is a deep learning model capable of considering both forward and backward contextual information in sequential data. By employing gating mechanisms such as input gates, forget gates, and output gates, BiLSTM successfully overcomes the challenges of gradient vanishing and gradient explosion often encountered by traditional Recurrent Neural Networks (RNNs) when processing long sequences. In the field of rolling bearing fault diagnosis, vibration signals possess distinct temporal characteristics. BiLSTM, through its forward and backward LSTM layers, can simultaneously capture both past and future information of the signal, thereby providing a more comprehensive learning of sequential features. The model structure is shown in Figure 3.

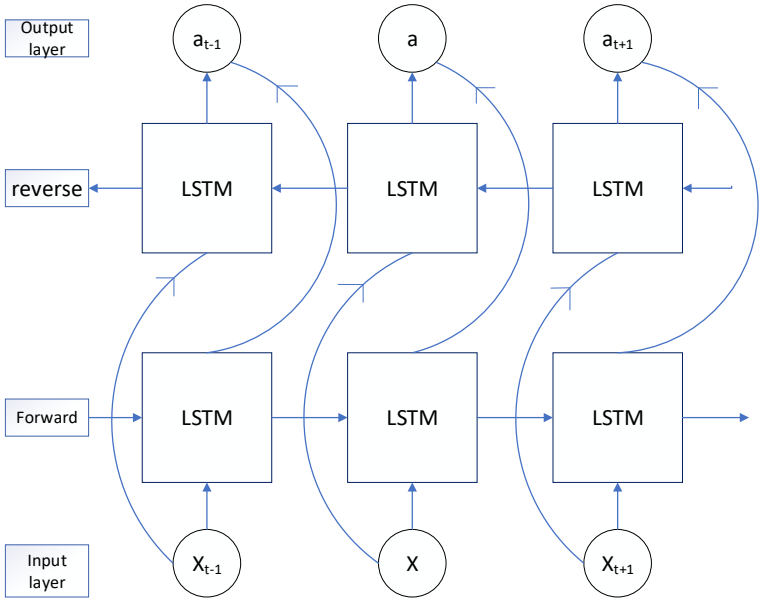

Figure 3. BiLSTM Model Diagram

Input Layer: This layer accepts the spatial features extracted by the Convolutional Neural Network (CNN) as the input sequence for the BiLSTM.

Forward LSTM Layer: This layer processes the input sequence in chronological order to capture the forward temporal features of the signal. It learns the dependencies and patterns from past to future within the sequence.

Backward LSTM Layer: This layer processes the input sequence in reverse chronological order to capture the backward temporal features of the signal. It learns the dependencies and patterns from future to past within the sequence.

Concatenation Layer: The outputs from the forward and backward LSTM layers are concatenated to form a comprehensive representation of the temporal features. This integration allows the model to leverage both past and future context simultaneously.

Output Layer: The output layer utilizes the Softmax function to output the probabilities of different fault categories, thereby completing the fault classification task.

The application of BiLSTM enables more accurate extraction of temporal features from vibration signals, thereby enhancing the performance of fault diagnosis.

2.3. Transformer Encoder Layer

The Transformer is a deep learning architecture that originated in the field of natural language processing. Its core innovation lies in the use of self-attention mechanisms to establish direct connections between different parts of the input sequence, thereby capturing global dependencies within the data. Compared to traditional temporal models, the Transformer demonstrates superior capability in capturing long-range dependencies and complex features through its multi-head attention mechanism. Additionally, it achieves higher parallel computing efficiency, which significantly enhances the model's feature extraction and representation abilities.

The model proposed in this paper employs a standard Transformer encoder architecture, with each encoder layer integrating multi-head self-attention mechanisms and a feed-forward neural network. Specifically, the multi-head self-attention mechanism consists of two attention heads, each with a dimension of 128. The feed-forward neural network has a hidden layer of 128 dimensions and employs the ReLU activation function.

During the preprocessing stage, feature data are processed through a series of convolutional and pooling layers to extract frequency-domain features from the CNN. These features are then fed into the Transformer encoder layer. Similarly, the temporal features processed by BiLSTM are also input into the Transformer encoder layer. Within the encoder layer, the global temporal dependencies and feature interactions of the signal are fully modeled.

By incorporating the Transformer encoder layer, the model is able to more comprehensively capture both global and local features within the vibration signals, enhancing the feature extraction and fusion capabilities, and thereby improving the accuracy and generalization ability of fault diagnosis.

2.4. Cross-Attention Mechanism (CrossAttention)

The Cross-Attention mechanism is an attention-based approach used to fuse features from different sources or of different types. Its fundamental idea is to compute attention weights between two distinct feature sequences, thereby effectively integrating them. Specifically, during the operation of the Cross-Attention mechanism, one feature sequence is regarded as the Query, while the other feature sequence is treated as both the Key and Value.

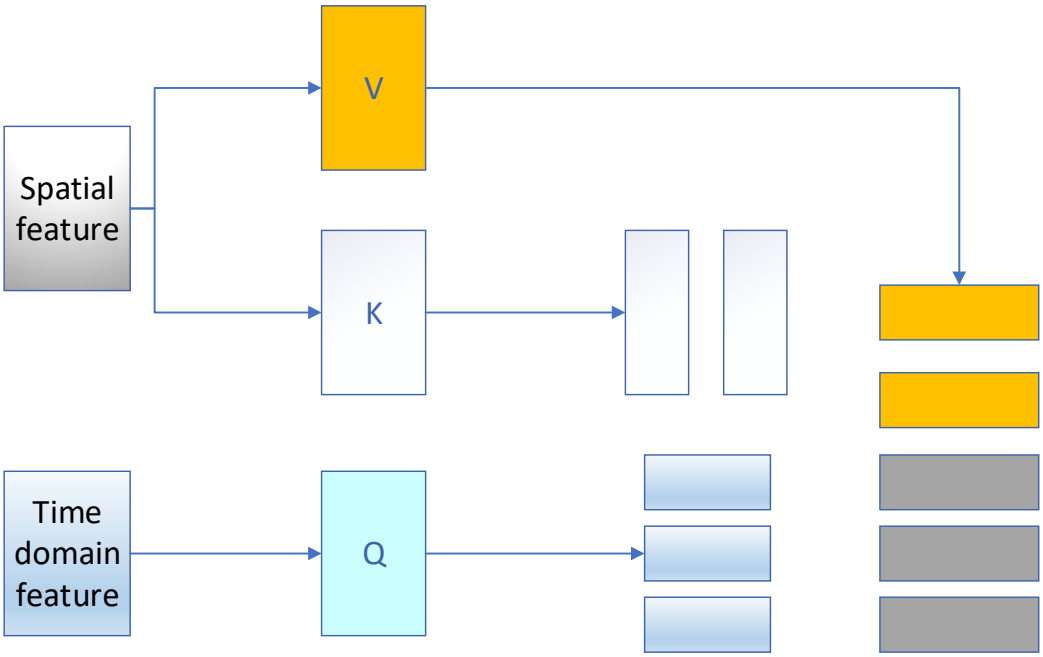

Figure 4. Cross-Attention Mechanism

During the fusion process, the spatial features extracted by the CNN-Transformer, denoted as Fspace, are used as the query sequence, while the temporal features extracted by the BiLSTM-Transformer, denoted as Ftime, are used as the key-value pair sequences. These features are first linearly mapped to the same dimension through learnable weight matrices WQ, WK, and WV, as shown in Equations (1), (2), and (3), respectively:

$$Q = W_Q \cdot F_{space} \quad (1)$$

In the equations: Q is the query sequence, dimensionless; W_Q is the weight matrix, dimensionless; F_{space} is the spatial feature parameter, dimensionless.

$$K = W_K \cdot F_{time} \quad (2)$$

In the equations: K is the key sequence, dimensionless; W_K is the weight matrix, dimensionless; F_{time} is the temporal feature parameter, dimensionless.

$$V = W_V \cdot F_{time} \quad (3)$$

In the equations: V is the value sequence, dimensionless; W_V is the weight matrix, dimensionless; F_{time} is the temporal feature parameter, dimensionless.

The attention weights are obtained by calculating the similarity between the query and the keys, and then these weights are used to perform a weighted sum of the values to obtain the fused feature representation, as shown in Equation (4):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

In the equations, dk represents the dimension of the key, which is dimensionless and used for scaling the dot product to prevent large values from causing gradient vanishing. The transpose of K (i.e., KT) helps avoid excessively large values that could lead to gradient vanishing.

In rolling bearing fault diagnosis, vibration signals contain rich time-frequency features. Features extracted by Fast Fourier Transform (FFT) and Variational Mode Decomposition (VMD) represent the frequency-domain and time-domain information of the signal, respectively. The Cross-Attention mechanism can effectively integrate these features from different domains. Specifically, the frequency-domain features extracted by FFT are used as the query, while the time-domain features extracted by VMD are used as the key and value. The attention weights are calculated between these two sets of features, thereby fusing the frequency-domain features with multi-scale time-domain features. This approach fully utilizes the diverse characteristics of the signal, thereby enhancing the accuracy and robustness of fault diagnosis.

3. Experimental Results and Analysis

3.1. Data Source

The dataset used in this experiment was obtained from the experimental testing system of Case Western Reserve University (CWRU) in the United States, as shown in Figure 5. This dataset has been widely used by scholars both domestically and internationally, making it highly authoritative and representative. For this experiment, we selected the drive-end bearing data with a sampling frequency of 12 kHz, a motor speed of 1797 r/min, and data collected at a rate of 12,000 samples per second. The bearing type used was SKF6205. The dataset includes three different fault modes: inner race fault, outer race fault, and rolling element fault. Each fault mode is further divided into three levels of severity: mild, moderate, and severe. Including the normal condition, there are a total of 10 data types.

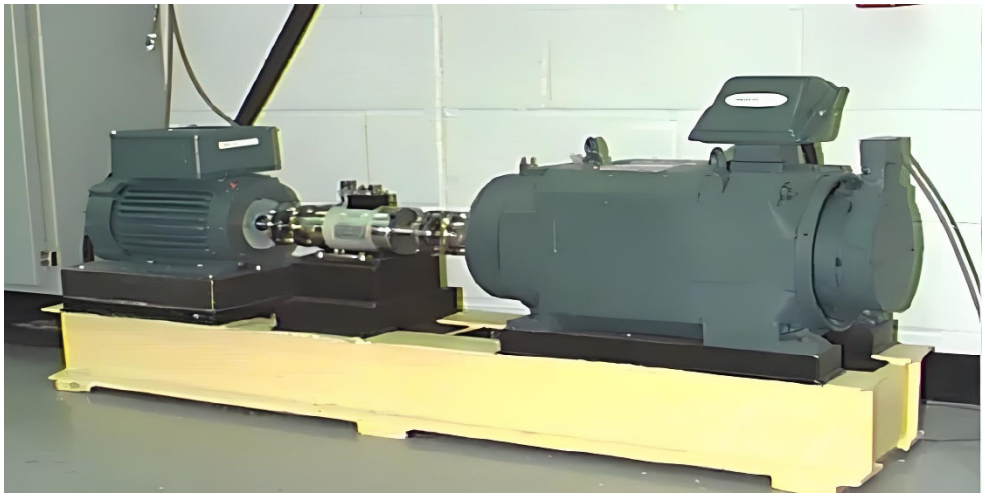

Figure 5. Case Western Reserve University Experimental Test Platform

The dataset can be segmented in various ways, and the segmentation stride also varies. In this experiment, the data segmentation was performed with fixed parameter settings. The stride was set to 512, with a sliding overlap rate of 0.5. The total number of fault samples was 2330, including 9 types of fault samples and 1 type of normal sample. After segmentation, the samples were divided into training, validation, and testing sets in a ratio of 7:2:1.

3.2. Experimental Results

After the one-dimensional fault signals were transformed and decomposed, they were fed into the model for training. The training results are shown in Figures 6, 7, 8, and 9.

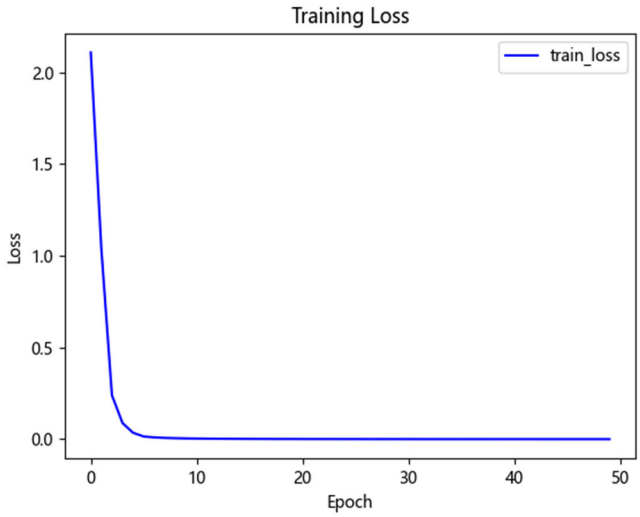

Figure 6. Training Loss

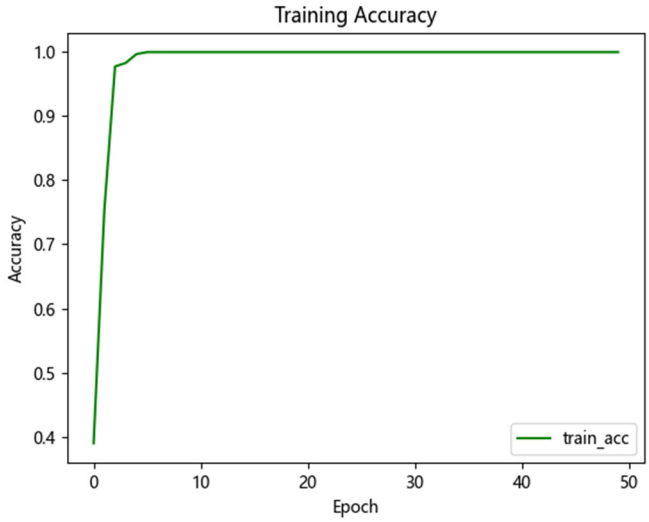

Figure 7. Training Accuracy

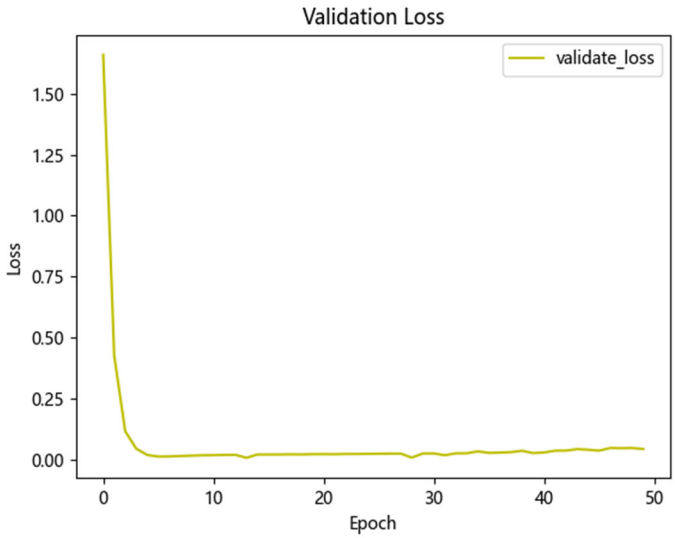

Figure 8. Validation Loss

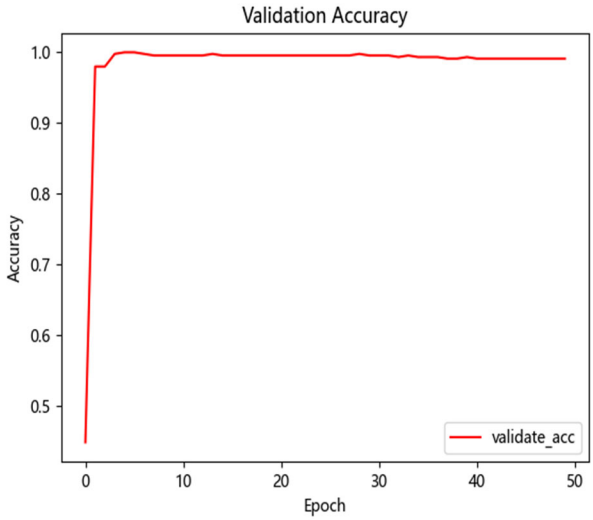

Figure 9. Validation Accuracy

Figures 6, 7, 8, and 9 illustrate the training loss, training accuracy, validation loss, and validation accuracy of the model across different epochs (training iterations). In these figures, the x-axis represents the number of training iterations, while the y-axis reflects the changes in loss values or accuracy. Observations reveal that as training progresses, the training loss rapidly decreases, while the validation loss and validation accuracy remain relatively stable.

Although the performance metrics on the training set provide feedback on the model's learning process, these indicators do not comprehensively reflect the model's generalization ability in real-world applications. To precisely evaluate the model's generalization performance, further assessment will be conducted on an independent test set. The test set, which was predefined and not used during model training, ensures the objectivity and reliability of the evaluation results. When assessing the performance on the test set, key metrics such as accuracy, recall, F1 score, and the confusion matrix are considered.

Table 1. Partial Performance Metrics Overview

precision	recall	f1-score	support
0	**1.0**	**1.0**	**27**
1	**1.0**	**1.0**	**28**
2	**1.0**	**1.0**	**24**
3	**1.0**	**1.0**	**20**
4	**1.0**	**1.0**	**28**
5	**1.0**	**1.0**	**13**
6	**1.0**	**1.0**	**22**
7	**1.0**	**1.0**	**21**
8	**1.0**	**1.0**	**22**
9	**1.0**	**1.0**	**19**

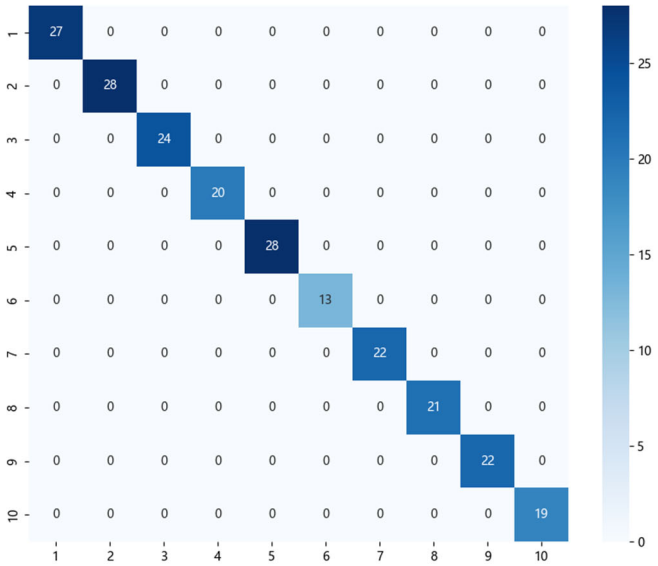

Figure 10. Confusion Matrix

The accuracy, recall, and F1 score of the model are shown in Table 1, all of which are 1.0000. The confusion matrix generated by running the model in this paper is shown in Figure 10. In this matrix, each row corresponds to the true class, while each column corresponds to the predicted class. The numbers within the matrix represent the counts of instances where the true class matches the predicted class at a specific position. For example, the number 19 in the first row and first column indicates that the model correctly predicted 19 instances when the actual class was A. Similarly, each cell in the matrix provides information on the model's accuracy in predicting different classes. The results show that the model achieved an accuracy of 100% in detecting each fault category.

3.3. Comparison with Different Models

To validate the efficiency and advantages of the proposed optimized CNN-BiLSTM-Transformer fault diagnosis model in terms of precision and execution speed, we compared it with the CNN-Transformer and CNN-BiLSTM fault diagnosis models using the same dataset as in this study. To further examine the generalization ability and robustness of the models and eliminate the influence of other random factors, we randomly executed the training process for each model 10 times. The comparative results of the models' performance after training are presented in Table 2.

Table 2. Comparison of Model Performance Metrics

Comparison Model	Average Accuracy	Average Runtime (s)
CNN-Transformer	**93.95**	**733.78**
CNN-BILSTM	**94.96**	**713.35**
CNN-BILSTM-Transformer	**99.99**	**634.17**

4. Conclusion

To achieve higher accuracy and faster operational efficiency in rolling bearing fault diagnosis, this paper proposes a fault diagnosis method that combines multiple neural networks with cross-attention mechanisms. The method employs Fast Fourier Transform (FFT) and Variational Mode Decomposition (VMD) to extract time-frequency domain features from the signals, enabling the extraction of multi-scale features from fault signals. Subsequently, the CNN-Transformer network is utilized to extract spatial features from the preprocessed multi-scale features of the fault signals, while the BiLSTM-Transformer network captures the temporal features. Finally, cross-attention is introduced to fuse these features, thereby enhancing the feature representation capability for fault signal recognition. The experimental results demonstrate that the proposed fault diagnosis method achieves an accuracy of 99% and outperforms other compared models in terms of operational efficiency.

References

[1] Guo Y ,Zhou J ,Dong Z , et al.Research on bearing fault diagnosis based on novel MRSVD-CWT and improved CNN-LSTM[J].Measurement Science and Technology,2024,35(9):

[2] Kang J ,Zhu X ,Shen L , et al.Fault diagnosis of a wave energy converter gearbox based on an Adam optimized CNN-LSTM algorithm[J].Renewable Energy,2024,231121022-121022.

[3] Xi Tao, Yang Weizhen. Research on Gearbox Fault Diagnosis Based on Optimized VMD and CNN [J]. Mechanical Science and Technology: 1-11

[4] Mao Chenglong, Zhang Mei, Zhang Jie. Rolling Bearing Fault Diagnosis Based on MSSA-VMD and CNN-BiLSTM [J]. Journal of Chongqing Technology and Business University (Natural Science Edition): 1-10

[5] Wang Minjuan, Jia Qian, Wang Youming, Ding Wenke. Rolling Bearing Fault Diagnosis Method Based on IMSE and Parameter-Optimized VMD [J]. Journal of Xi'an University of Posts and Telecommunications: 1-10.

[6] Cao Jingsheng, Yu Yang, Wang Qi, Dong Yining. Intelligent Fault Diagnosis of Motor Bearings Based on Optimized VMD-CNN-BiLSTM [J]. Modern Electronics Technique, 2024, 47(12): 115-121.

[7] Xie Fengyun, Wang Gan, Shang Jiandong, Fan Qiuyang, Zhu Haiyan. Gearbox Fault Diagnosis Based on Adaptive Variational Mode Decomposition [J]. Journal of Propulsion Technology: 1-11.

[8] Meng Jingyu, Yang Liming, Zhang Cheng, Wu Boyang, Xu Guoping, Yu Jian. Feature Extraction and Diagnosis of Wind Turbine Gearbox Faults Based on Zoom-FFT-CEEMD and Wavelet Packet Denoising [J]. Micro Motors, 2024, 52(04): 28-32=37.

[9] Gao Hongwei, Li Xincheng, He Xiaoning, Liu Chang'an, Wang Hanchi, Li Wencheng, Zhang Baohui. Research on Wind Turbine Gearbox Fault Diagnosis Method Based on Hilbert Transform [J]. Machine Tool & Hydraulics, 2024, 52(09): 215-220.

[10] Wei Hangxin, Cheng Huan, Wu Wei, Wang Xiaorong. Rolling Bearing Fault Diagnosis Based on VMD Visualization and Deep Learning [J]. Mechanical Design and Manufacture: 1-6.

[11] Zhu Junjie, Zhang Qinghua, Zhu Guanhua, Su Naiquan. Rolling Bearing Fault Diagnosis Based on EEMD and CNN-SVM [J]. Machine Tool & Hydraulics: 1-9.

[12] Zhou Ruifeng. Research and Implementation of Intelligent Fault Diagnosis Methods for Rolling Bearings [D]. Dalian University of Technology, 2009.

[13] Xu Jingwen, Yang Ping, Yin Xiaojun. Gear Fault Diagnosis Based on EMD Decomposition and Levy-SSA-BP Neural Network [J]. Mechanical Transmission, 2024, 48(05): 152-157.

[14] Long Xiafei, He Zhicheng, Zhou Ling, Liu Weiqiang, Liang Kai. Research on Wind Turbine Gearbox Fault Diagnosis Based on KOA-CNN-BiLSTM-AM [J]. Machine Tool & Hydraulics: 1-10.

[15] Wang Jinhua, Liu Zhengqi, Cao Jie, Liu Yunqiang, Chen Li. Gearbox Fault Diagnosis Based on R Vine Copula-DBN [J]. Journal of Beijing University of Aeronautics and Astronautics: 1-15.

[16] Wang Chengbing. Gearbox Fault Diagnosis Based on Multi-Sensor Information Fusion and Dual-Stream CNN [J]. Construction Machinery, 2024, 55(04): 69-78+10-11.

[17] Xu M ,Yu Q ,Chen S , et al.Rolling Bearing Fault Diagnosis Based on CNN-LSTM with FFT and SVD [J]. Information, 2024, 15(7):399-399.

[18] Yu T ,Li C .Fault diagnosis method of rolling bearing based on MSSA-VMD algorithm[J].Journal of Physics: Conference Series, 2024,2752(1):

[19] Cui K ,Liu M ,Meng Y .A new fault diagnosis of rolling bearing on FFT image coding and L-CNN[J].Measurement Science and Technology,2024,35(7):

[20] Zhenzhen J ,Deqiang H ,Zexian W .Intelligent fault diagnosis of train axle box bearing based on parameter optimization VMD and improved DBN[J].Engineering Applications of Artificial Intelligence,2022,110.