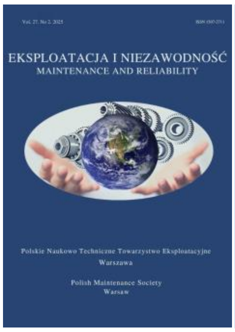


**Eksploatacja i Niezawodnosc – Maintenance and Reliability**

**Volume 27 (2025), Issue 2**

journal homepage: http://www.ein.org.pl

Article citation info:
Wei L, Peng X, Cao Y, Multi-Scale Graph Transformer for Rolling Bearing Fault Diagnosis, Eksploatacja i Niezawodnosc – Maintenance and Reliability 2025: 27(2) http://doi.org/10.17531/ein/194779

# Multi-Scale Graph Transformer for Rolling Bearing Fault Diagnosis

Indexed by:

**Lunpan Wei[a], Xiuyan Peng[a], Yunpeng Cao[b,*]**

[a] College of Intelligent Systems Science and Engineering, Harbin Engineering University, China
[b] College of Power and Energy Engineering, Harbin Engineering University, China

## Highlights

- Multi-Scale Graph Transformer enhances fault diagnosis precision significantly.
- Graph Node Aggregation Mechanism broadens receptive fields for feature extraction.
- Centrality and Spatial Encoding capture intricate graph node structural insights.
- Transformer Self-Attention improves crucial fault signature identification.

## Abstract

Traditional graph neural networks commonly face limitations in fault diagnosis due to insufficient feature extraction at a single scale, particularly in complex operational scenarios. To address these challenges, we introduce an innovative multi-scale graph Transformer framework for rolling bearing fault diagnosis. This framework incorporates a unique multi-scale feature aggregation mechanism, along with centrality and spatial encoding of graph nodes, to enhance structural insights. By leveraging multi-head self-attention, our approach efficiently extracts and learns fault features, thereby significantly improving fault identification. Extensive experiments on the designed bearing dataset, along with a customized rolling bearing apparatus, validate the efficacy of our method. Our model achieves a peak diagnostic precision of 99.5% and maintains an average accuracy of over 97.9%, underscoring its robustness and adaptability across diverse operational scenarios.

## Keywords

multi-scale graph transformer, rolling bearings, advanced fault diagnosis, feature extraction, deep learning

## 1. Introduction

Rolling bearings, as indispensable components in rotating machinery, are referred to as the "joints of industry." Operating in extremely complex environments, they are prone to damage. With increasing operational time, rolling bearings are more susceptible to failure. If these faults are not detected and addressed in a timely manner, they can significantly reduce the safety and lifespan of mechanical equipment. Traditional methods for diagnosing bearing faults typically involve signal decomposition techniques for feature extraction, such as Fourier transform (FT), empirical mode decomposition (EMD), and singular value decomposition (SVD) 123. Subsequently, classification is performed using methods like support vector machines (SVM) and Bayesian classifiers 45. However, while traditional methods perform well with small-scale datasets, they have limited feature extraction capabilities when faced with large-scale datasets, making it difficult to fully exploit the potential relationships among fault data 6.

As computing hardware such as graphics processing units (GPUs) advances rapidly, deep learning approaches for intelligent fault diagnosis have flourished. Among the various

(*) Corresponding author.
E-mail addresses: L. Wei (ORCID: 0000-0003-2416-6630) weilunpan@hrbeu.edu.cn, X. Peng pengxiuyan@hrbeu.edu.cn,
Y. Cao (ORCID: 0000-0002-4953-8886) caoyunpeng@hrbeu.edu.cn,

deep neural network (DNN)-based methods, including deep belief networks (DBNs), autoencoders (AEs), convolutional neural networks (CNNs), and long short-term memory networks (LSTMs), there has been a surge of interest from academic circles and industry alike 78. Ma and his team have devised a novel predictive model that integrates particle filter and LSTM neural networks to forecast the remaining service life of rotating equipment. This hybrid approach addresses the shortcomings of traditional data-driven techniques, resulting in improved prediction accuracy 9. Shao and his colleagues investigated the use of a convolutional deep belief network (CDBN) for diagnosing bearing faults. To enhance the network's performance, they introduced exponential moving average methods, effectively leveraging historical data trends 10. Wang and his research team introduced an innovative approach that combines multi-sensor data fusion and a CNN optimized with a bottleneck layer for rotating machinery fault recognition. This methodology mitigates the limitations of single-sensor features and enhances the feature set, enabling more accurate fault detection 1112.

In bearing fault diagnosis, the relationships between fault signals vary significantly with changes in machine health status. Effectively modeling and analyzing the intricate relationships between signals play a crucial role in accurately diagnosing machine faults. DNNs can effectively learn the correlations between input features. However, in the self-learning process of feature representation, accurately uncovering the relationships among fault data is challenging 1314.

Graph data, based in non-Euclidean spaces and including information about nodes and edges between nodes, can accurately uncover the complex relationships among fault data, providing a more comprehensive information representation capability 1516.

Graph Neural Networks (GNNs) have demonstrated outstanding performance in handling non-Euclidean data, as in recommendation systems, link prediction, node classification, and protein structure inference 1718. They can fully exploit the relationships between nodes for feature extraction, presenting new opportunities for the development of rolling bearing fault diagnosis.

While GNNs excel in non-Euclidean data, they are limited to aggregate information from a fixed number of nodes around the fault node, thus failing to fully extract fault features. Wang et al. addressed this limitation by modeling data through spatiotemporal graphs, effectively utilizing temporal and spatial information to enhance the model generalization 19. Combining graph convolutional networks with empirical mode decomposition, Hong et al. were able to mine deep features of fault signals 20. However, these methods did not focus on useful features. Feng et al. improved the feature selection capability by employing multi-head attention in Transformers to focus on more important features 21. Liu et al. combined wavelet time-frequency analysis with the Swin Transformer to enhance fault diagnosis capability using its powerful image classification abilities 22. Huang et al. constructed feature extractors using a variant of Transformer, VOLO, to obtain finer-grained fault feature representations 23. Nonetheless, these methods still face several challenges: 1) Weakness in single-scale feature extraction, insufficiently capturing fault node features; 2) Transformer-based models overlook the complex relationships among fault data, limiting feature representation capability; 3) Current methods are ineffective in handling noise and variations in operating conditions during fault diagnosis in complex environments.

To address the aforementioned issues, this paper proposes the following solutions:

1. Weak Single-Scale Feature Extraction: To address the weak single-scale feature extraction, our method employs Chebyshev graph convolutional networks with varying scales of receptive fields. This multi-scale approach allows aggregation of information from different neighborhood ranges, enhancing the model's ability to capture comprehensive fault features. By extracting features at multiple scales, our model identifies subtle differences in fault characteristics, which result in more accurate fault diagnosis.

2. Overlooking Complex Relationships Among Fault Data: Our proposed Multi-Scale Graph Transformer (MSGraphormer) leverages graph attention mechanisms within the Transformer framework to adaptively learn the significance of various fault features. This enables the model to focus on critical features and understand complex relationships within the fault data, thereby improving overall feature representation. By incorporating centrality encoding and spatial encoding for graph nodes, our method enhances structural insights and captures intricate graph

node relationships more effectively.

3. Ineffectiveness in Handling Noise and Variations in Operating Conditions: To improve noise resistance and robustness, our method incorporates Gaussian white noise into the datasets and evaluates performance across different signal-to-noise ratios (SNRs). The multi-scale and Transformer-based architecture allows our model to maintain high accuracy even in noisy conditions. Our experimental results show that our model demonstrates strong noise resistance, ensuring reliable fault diagnosis in real-world scenarios with varying noise levels and operating conditions.

By addressing these critical issues, MSGraphormer significantly enhances the accuracy, robustness, and reliability of rolling bearing fault diagnosis in complex environments. Therefore, this paper proposes a method of multi-scale graph Transformer. This method constructs a novel graph node feature aggregation model, aggregating feature information from each graph node neighborhood into the feature representation of the central node. It can fuse multi-scale node information from different neighborhoods to enhance feature representation, reducing interference from noise, variations in operating conditions, and other complex environmental factors. By encoding fault structure information into Transformers through the centrality encoding and spatial encoding of graph nodes, the complex relationships among fault signals can be captured. Experimental results on rolling bearing fault diagnosis demonstrate the accurate classification of bearing faults under different operating conditions.

The remainder of the paper is structured as follows: In Section 2, we review the related work. Section 3 outlines the Multi-scale Transformer Model used in our research. In Section 4, we present our experimental results. Finally, Section 5 concludes the paper with a summary of our contributions and suggestions for future work.

## 2. Related Works

This section reviews advancements in Graph Neural Networks (GNNs) and structure encoding techniques, emphasizing their applications in fault diagnosis and structural information extraction to enhance diagnostic accuracy.

### 2.1. Graph Neural Networks

GNN are a deep learning algorithms capable of effectively handling non-Euclidean data 24. Graph Convolutional Networks (GCNs) are a specialized form of convolutional neural networks designed to operate on graphs. GCN convolutions can be realized in two distinct ways: spectral-based convolutions and spatial-based convolutions. Spectral-domain GCNs employ the spectral decomposition of the graph Laplacian to capture signal correlations. Specifically, the graph convolution is formulated as:

$$f * x = U((U^T f) \odot (U^T x)) = U g_w U^T x \qquad (1)$$

Where, $f \in \mathbb{R}^N$ and $x \in \mathbb{R}^N$ represent the convolutional kernel and feature vector, respectively. $U$ is the Fourier basis, derived from the eigendecomposition of the Laplacian matrix $L = U\Lambda U^{-1}$ , where $U$ and $\Lambda$ are the eigenvector and eigenvalue matrices of $L$ . $g_w = \mathrm{diag}(U^T h) = \mathrm{diag}(w_1, w_2, \dots, w_N)$ denotes a set of learnable parameters.

Due to the high computational cost of directly performing eigendecomposition on graphs, Kipf and colleagues proposed approximating $g_w$ with a Chebyshev polynomial:

$$g_w = \sum^{\substack{k=0\\K}} w_k T_k(\hat{\Lambda}) \qquad (2)$$

Where, $\hat{\Lambda} = \Lambda - I_n$ with $I_n$ being the identity matrix. By limiting the approximation to the first-order (K=1) and utilizing Chebyshev polynomial properties, the graph convolution can be simplified to:

$$\begin{aligned} f^* x &= U \sum^{\substack{k=0\\K}} w_k T_k(\bar{\Lambda}) \cdot U^T x \\ &= \sum^{\substack{k=0\\1}} w_k T_k(U\Lambda^{\sim} U^T)x = w_0 + w_1(L - I_n)x \end{aligned} \qquad (3)$$

By setting $w^{'} = w_0 = -w_1$ and further simplifications, the standard GCN forward propagation equation emerges:

$$Z = \sigma(D^{-0.5} A D^{-0.5}(XW)) \qquad (4)$$

Where, $\hat{A} = A + I_n$, $A$ is the adjacency matrix of $X$, $\widehat{D} = \sum_j \hat{A}_{ij}$, $W$ is the parameter matrix to be learned, and $\sigma$ is the activation function. The standard GCN structure is illustrated in Figure 1.

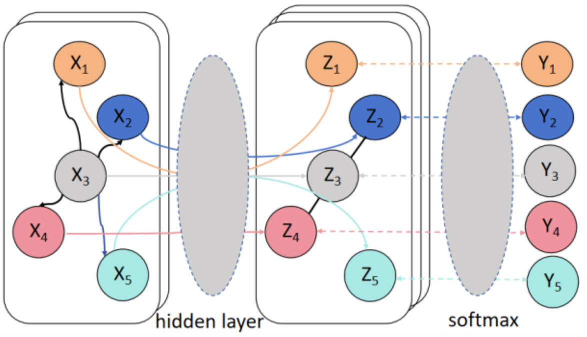


Figure 1. Standard Graph Convolutional Network Architecture.

### 2.2. Structure Encoding

In addition to the inherent feature information of nodes, their structural information also provides valuable content within the graph. In this paper, two methods, namely centrality encoding and spatial encoding, are employed to embed the structural information of graph nodes 25.

1. Centrality Encoding

Centrality information, based on the degree of each node, is embedded and can be directly incorporated into the node's features as input to the Transformer encoder 26.

$$h_i^{(0)} = x_i + z_{deg(v_i)} \tag{5}$$

2. Spatial Encoding

For sequential data, there are two ways to embed spatial information: absolute position encoding and relative position encoding. However, for graph data, nodes are not arranged in order but connected through edges 27. To encode the structural information of graph data, this paper proposes the following method: $\phi(v_i, v_j)$. In the context of graph $G$, when two nodes $v_i$ and $v_j$ are directly connected by an edge, indicating mutual adjacency, the spatial encoding for these nodes is chosen to be the shortest path between them, which in this scenario is simply the direct edge itself. However, if there is no direct edge connecting $v_i$ and $v_j$, a default or placeholder value of $\phi$ is assigned, typically set to -1, to indicate the absence of a direct connection. The embedding process proceeds accordingly by applying this logic to determine the spatial encoding for each pair of nodes in the graph.

$$A_{ij} = (h_i W_Q)(h_j W_K)^T / \sqrt{d} + b_\phi(v_i, v_j) \tag{6}$$

where, $b_\phi(v_i, v_j)$ is a learnable parameter that is shared across all layers.

After obtaining the feature matrix encoded with embeddings of two types of structures, it is then input into a multi-scale feature aggregation module to fuse information from neighboring nodes of different scales.

## 3. Multi-scale Transformer Model

The multi-scale Graph Transformer model introduces a novel graph node feature aggregation module that effectively leverages diverse neighborhood feature information to broaden the receptive field of nodes 28. By aggregating feature information from neighboring nodes at different scales, it enhances the feature representation of central nodes. Additionally, incorporating centrality encoding and spatial encoding enriches node feature information by embedding the positional information of graph nodes.

### 3.1. Multi-scale Feature Aggregation Module

Within the framework of graph neural networks (GNNs), incorporating information from neighboring nodes into node feature representations plays a pivotal role. Figure 2 illustrates a schematic representation of the multi-level feature integration mechanism.

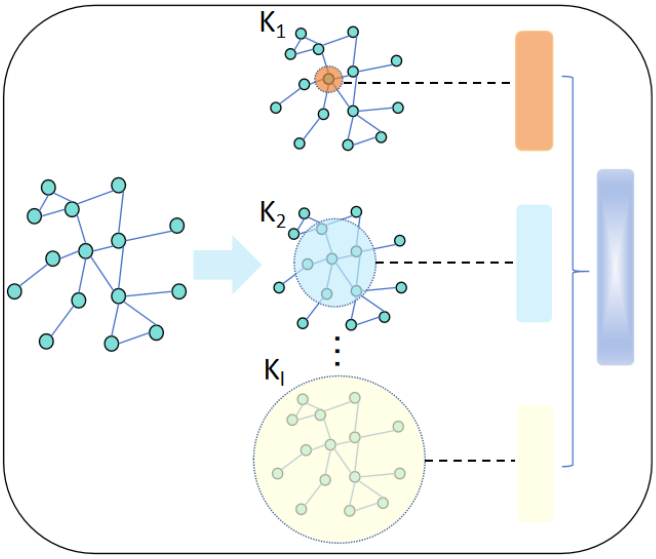


Figure 2. Multi-level Feature Integration Mechanism.

To enhance feature representation, this paper integrates information from different neighborhoods to design the multi-scale feature aggregation module. For a node $v$, let $N^k(v) = \{u \in v \mid d(v,u) \leq k\}$ denote its k-th order neighborhood, where $d(v,u)$ represents the shortest path distance between $v$ and $u$. $N^0(v) = \{v\}$ is defined as the node itself.

In the multi-scale aggregation module, $N^k(v)$ is defined as the aggregation of neighborhood feature information for node $v$, denoted as:

$$x_v^k = N^k(v) \tag{10}$$

The neighborhood feature information $S_v = (x_0^v, x_1^v, \dots, x_k^v)$ of the target node can be obtained using Equation (5), where $k$ is fixed as a hyperparameter. Thus, for each node in graph $G$, its corresponding neighborhood feature information can be generated, denoted as $X_G \in \mathbb{R}^{n \times d \times (k+1)}$. Each node in the graph aggregates feature information from its k scales of neighborhood nodes. Consequently, the feature matrix of nodes in graph $G$ is expanded to $S = (X_k^0, X_k^1, \dots, X_k^n)$, where $X_k^n \in \mathbb{R}^{d \times (k+1)}$represents the aggregation of feature information from

$k$ -th order neighborhood nodes for the initial feature matrix $X_0^n$.

As previously mentioned, the receptive field is contingent upon the value of $k$. Consider a normalized adjacency matrix $\hat{A}$ and a feature matrix $X$. In the scenario where $k = 0$, the aggregated node features are derived directly from the nodes themselves, specifically the matrix $X$. Conversely, when $k$ equals 1, the aggregated features are calculated as the average of the immediate neighboring nodes, represented by the matrix multiplication $\hat{A}X$. For higher values of $k$, such as $k = 2$, the aggregated node features encompass both the first-order and second-order neighbors, achieved through consecutive matrix multiplications like $\hat{A}\hat{A}X$. This trend continues for higher orders of $k$.

Thus, the $k$ -th order neighborhood node feature matrix can be represented as:

$$X_k = \hat{A}^k X \tag{11}$$

Algorithm 1 provides a specific implementation method. Traditional GNN architectures only consider a single-scale information aggregation method, where a fixed $k$ is used to aggregate information from neighboring nodes. Therefore, this paper introduces a multi-scale feature aggregation module aimed at integrating different $k$ values, representing various scales of neighborhood node feature information, to enhance the feature representation of nodes.

**Algorithm 1:** Multi-scale Feature Aggregation

```
Input: Standardized adjacency matrix $A$, feature
       matrix $X$, aggregation scale $K$
Output: Node feature vectors $XG$
$Fork \leftarrow 0$;
for $k \leftarrow 0$ to $K$ do
    for $i \leftarrow 0$ to $n$ do
        $Xg_{i,k} \leftarrow X_i$
    end
    $X \leftarrow A \times X$;
end
return $XG$
```

### 3.2. Overall Network Architecture

In this paper, a multi-scale graph transformer (MSGraphormer) is proposed based on a graph transformer for bearing fault diagnosis 29. This method effectively addresses the issues of insufficient structural exploration and feature extraction among fault vibration signal nodes. The overall architecture of the proposed rolling bearing fault diagnosis method is shown in Figure 3.

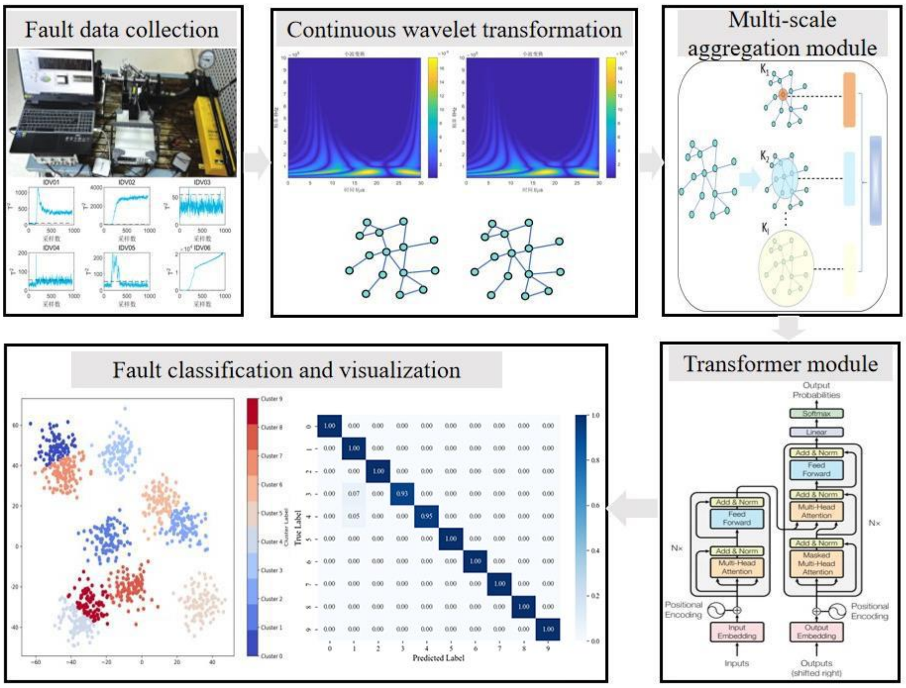


Figure 3. MSGraphormer Fault Diagnosis Model.

First, wavelet transform is applied to the collected raw bearing fault data to obtain the fault time-frequency map. Next, by calculating the distances between two fault feature representation nodes and selecting the k nearest ones (excluding the node itself), an adjacency matrix is constructed. This adjacency matrix, along with the feature representation of the fault nodes, are used as the input for the fault diagnosis model, MSGraphormer.

Within MSGraphormer, the feature matrix and adjacency matrix of the fault nodes are first passed through a multi-scale feature aggregation module to aggregate the k-scale neighborhood information of the fault nodes, denoted as $S = (X_{k_0}, X_{k_1}, \ldots, X_{k_n})$. Subsequently, a linear layer is employed to obtain a low-dimensional embedding representation of the fault nodes.

Subsequently, a linear layer is employed to obtain a low-dimensional embedding representation of the fault nodes.

$$Z_n^{(0)} = X_n^k W_E \tag{7}$$

Where $W_E \in \mathbb{R}^{d_{in} \times d_{out}}$, assuming the feature matrix dimension of the fault nodes is  and the output dimension of the Transformer embedding layer is $d_{out}$.

Then, the resulting tensor is fed into the encoder of the Transformer, which consists of a multi-head self-attention (SA) mechanism and a feed-forward network (FFN). Layer normalization (LN) is applied before prior to each layer. The FFN layer comprisesis constituted of two linear layers interspersed with a nonlinear activation function, GELU.

$$Z_i^{\prime(l)} = \mathbf{MSA}(\mathbf{LN}(Z_i^{(l-1)})) + Z_i^{(l-1)}) \tag{8}$$

$$Z_i^{(l)} = \mathrm{FFN}(\mathrm{LN}(z_i^{\prime(l)}) + Z_i^{(l-1)}) \tag{9}$$

After feature extraction by the Transformer, the characteristics of the faulty nodes are reorganized through a fully connected layer, and then accurately classified by a Softmax classifier.

## 4. Experimental Simulation and Result Analysis

To substantiate the proficiency of the novel approach in diagnosing rolling bearing faults, this investigation incorporated tests utilizing two distinct data sources: one from a benchmark dataset originating from the motor bearings of Case Western Reserve University (CWRU) in the United States, and the other from an experimental platform tailored specifically for rolling bearings. The deep learning framework selected for this research was PyTorch, and the experiments were executed on a Windows 10-based computer system, augmented by an AMD R5-3600 processor and a GTX3060 graphics processing unit (GPU), ensuring the computational requirements were met with efficiency.

### 4.1. Experiment with the CWRU Bearing Dataset

1. Experimental Dataset and Parameter Settings

The experiment used the open dataset from Case Western Reserve University for validation. The bearing fault testbed mainly consisted of an induction motor, torque sensor, and dynamometer. The experimental setup is shown in Figure 4. The deep groove ball bearings SKF6205 and SKF6203 were the subjects of the study. Vibration signals were collected under four different operating conditions, namely normal, inner race fault, outer race fault, and rolling element fault. Vibration signals for four fault levels were collected for rolling elements, while three fault levels were collected for the outer race. All vibration signals were collected under loads of 0, 1, 2, and 3 hp (1 hp≈ 0.735 kW) on the electric motor. Sampling frequencies were 12 and 48 kHz, and bearing data was obtained by sampling on the drive-end bearing or fan-end bearing. Normal condition referred to the absence of any fault.

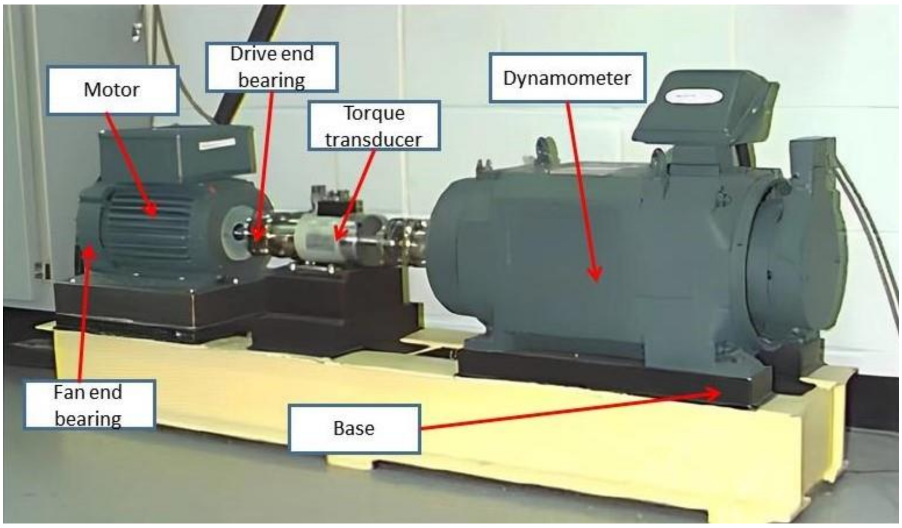
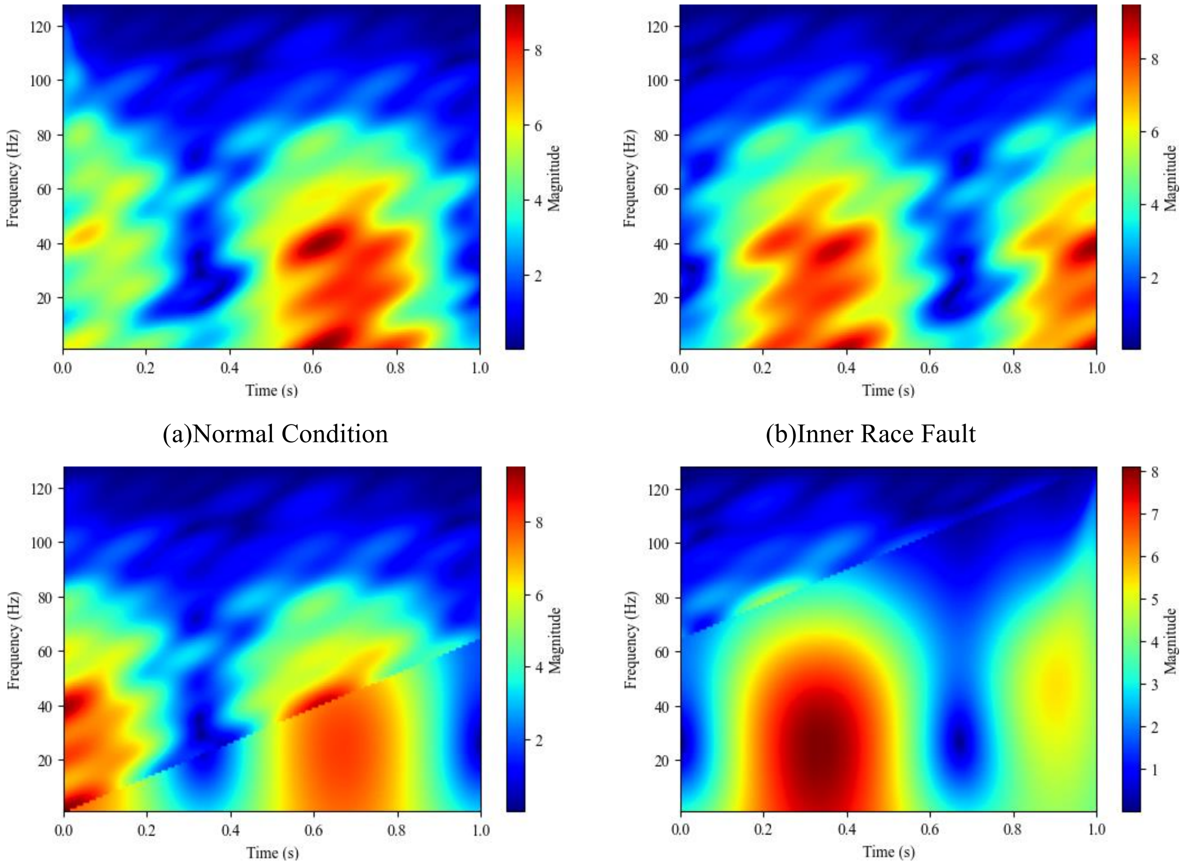


Figure 4. Experimental Device Platform.

To extract the time-frequency characteristics of the vibration signals, the original vibration time series was transformed into time-frequency maps through wavelet transform. When selecting data to generate time-frequency maps, the collected data was first standardized. After standardization, a sliding window was used to assist in sampling, with 1024 sampling points as a unit. The sliding window sampled the data in sequence with a step size of 128. The time-frequency maps generated by the wavelet transform of the original vibration time series were sized 32×32, as shown in Figure 5. The time-frequency maps possessed features from both the time domain and the frequency domain, amplifying relevant details of time and frequency, respectively, enabling the model to extract fault signal features more effectively.



(a)Normal Condition

(b)Inner Race Fault

(c)Rolling Element Fault

(d)Outer Race Fault

Figure 5. Time-Frequency Representations after Wavelet Transform.

To simulate real-world scenarios, the collected data in this paper was divided into 10 categories based on fault types, with each fault type containing 100 time-frequency map samples. Sixty percent of the data was designated as the training set, 20% as the test set, and the remaining 20% as the validation set. The breakdown of the experimental dataset is shown in Table 1.

Table 1. Experimental Dataset Breakdown.

| Fault Size (inches) | Label | Fault Type |
|---|---|---|
| 0.007 | 0 | Inner Ring Fault |
|  | 1 | Ball Fault |
|  | 2 | Outer Ring Fault |
| 0.014 | 3 | Inner Ring Fault |
|  | 4 | Ball Fault |
|  | 5 | Outer Ring Fault |
| 0.028 | 6 | Inner Ring Fault |
|  | 7 | Ball Fault |
|  | 8 | Outer Ring Fault |
|  | 9 | Normal State |

The experiment sets the number of layers of the Transformer to 5, the embedding layer dimension to 512, the initial learning rate to 0.0001, and the optimizer to Adam $W$. The iteration batch count threshold count is set to 100. To prevent overfitting, Dropout is set to 0.5.

2. Experimental Results and Performance Analysis

To validate the efficacy and advantage of our proposed approach, we constructed six distinct fault diagnosis models for comparison. These encompassed the standard Graph

Convolutional Network (GCN), as well as the Chebyshev Graph Convolutional Network (Cheby Net) that incorporated receptive fields spanning scales of 1, 2, and 3. Furthermore, we benchmarked our method against the traditional 5-layer Convolutional Neural Network (CNN) and Graphormer, a cutting-edge graph Transformer model.

To mitigate the impact of randomness on our findings, we conducted five independent experimental runs, thereby ensuring the robustness of our proposed methodology. The primary evaluation criterion was overall classification accuracy, encompassing the highest, lowest, and mean accuracies over the five iterations. The outcomes of these experiments are depicted in Figure 6 and summarized in Table 2.

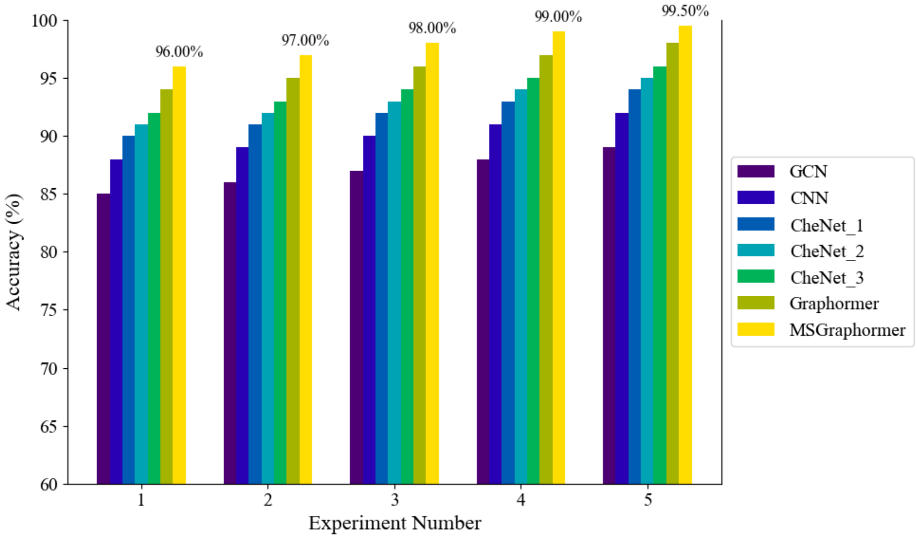


Figure 6. The Diagnostic Results of Different Methods in the CWRU Dataset Experiment.

Table 2. Diagnostic Results on the CWRU Dataset (%)

| Method | Max Accuracy | Min Accuracy | Average Accuracy ± Std Dev |
|---|---|---|---|
| GCN | 82.53 | 85.35 | 80.98 ± 1.64 |
| CNN | 87.35 | 89.35 | 88.49 ± 1.06 |
| Cheby Net_1 | 92.46 | 88.67 | 91.15 ± 0.81 |
| Cheby Net_2 | 90.78 | 92.75 | 91.57 ± 0.41 |
| Cheby Net_3 | 91.13 | 92.83 | 91.84 ± 0.86 |
| Graphormer | 92.33 | 95.13 | 93.24 ± 0.45 |
| MSGraphormer | 99.50 | 96.00 | 97.90 ± 0.10 |

Based on the results presented in Figure 6 and Table 2, it is evident that MSGraphormer outperforms the other seven methods in terms of stability. In each of the five trials, MSGraphormer achieved accuracies of 96.00%, 97.00%, 98.00%, 99.00%, and 99.50%, respectively. Furthermore, its average accuracy stands at 97.90% with a minimal standard deviation of 0.10%, highlighting the remarkable stability of the proposed approach. Although MSGraphormer demonstrates higher accuracy and stability, it only shows a 4.66% improvement in average accuracy compared to Graphormer. In the future, it will be necessary to further analyze differences in feature extraction and computational efficiency.

As observed in Table 2, the Chebyshev graph convolutional network (ChebyNet) exhibits a positive correlation between the order of the Chebyshev polynomial and its accuracy. This indicates that an larger receptive field scale leads to enhanced accuracy in graph convolutional networks (GCN).

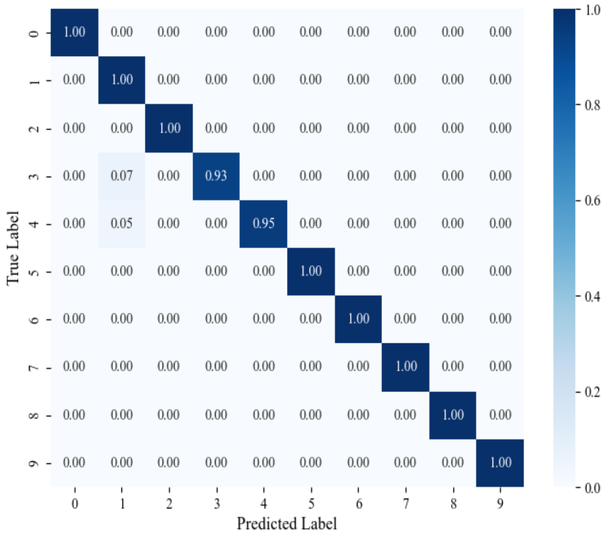


Figure 7. Confusion Matrix of Test Results from the Dataset Experiment.

MSGraphormer, incorporating multiple scales of receptive fields, is capable of not only aggregating information from neighboring nodes with varying receptive fields but also integrating this information to produce more discriminative features. Consequently, MSGraphormer demonstrates superior classification performance. The confusion matrix graph, as shown in Figure 7, provides a more intuitive demonstration of the classification results.

From Figure 7, the confusion matrix demonstrates the model's high performance, with values concentrated along the diagonal, indicating accurate predictions that matching true labels. For example, the values at positions (0,0), (1,1), and (2,2) are all 1.00, showing perfect accuracy for these labels. Some misclassifications are noted, such as 0.07 at position (3,1) and 0.05 at position (4,1), indicating that true labels 3 and 4 were occasionally misclassified as label 1. Overall, the high diagonal values and minimal off-diagonal values highlight the model's high accuracy and strong classification ability with minimal errors. To further highlight the superiority of the method proposed in this paper, t-SNE visualization results are presented in Figure 8.

As depicted in Figures 7 and 8, the approach presented in this paper demonstrates remarkable proficiency in classifying ten distinct fault types. This proficiency stems primarily from the methodology's ability to extract fault feature information at multiple scales, thereby augmenting the descriptive power of the features and enhancing the utilization of pertinent information. Moreover, the incorporation of Transformer technology enables the model to adaptively learn the representational capabilities of fault node features, further improving the overall performance of the classification system.

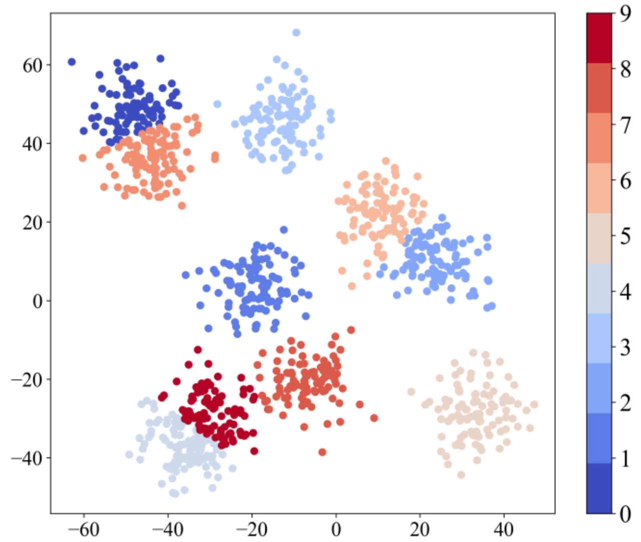


Figure 8. t-SNE Result Visualization.

### 4.2. Rolling Bearing Experiment on Experimental Platform

To further verify the feasibility and effectiveness of MSGraphormer, experiments were conducted on a rolling bearing experimental platform. The rolling bearing experimental platform, as shown in Figure 9, consists of vibration sensors, laser displacement sensors, and noise sensors, which are used to collect vibration, displacement, and noise fault signals, respectively.

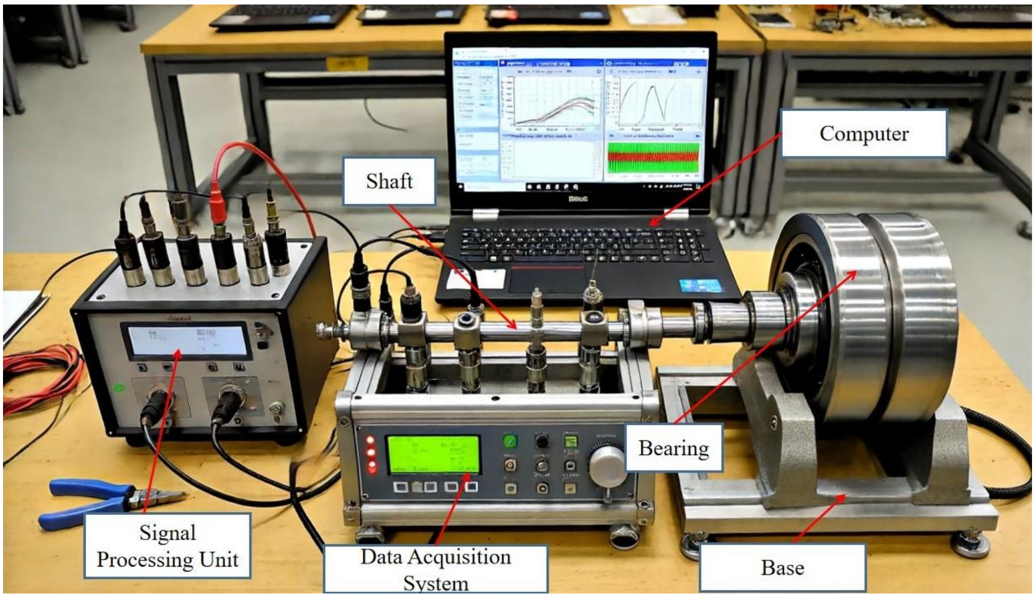


Figure 9. Rolling Bearing Test Bed.

The experimental platform operates at speeds of 1200, 1400, 1600, and 1800 rpm, with sampling frequencies of 12 and 36 kHz. It simulates four classic bearing fault conditions: normal condition, inner ring fault, rolling element fault, and compound fault (inner ring plus rolling element). The platform consists of several key components. The Signal Processing Unit is connected to multiple sensors that measure various parameters. The central component is the Shaft, which is supported by a Bearing. Data from the sensors is collected and processed by the Data Acquisition System, and results are displayed on a computer. The entire assembly is mounted on a stable Base to ensure accurate measurements and reliable operation.

During the experiment, a radial force of 100N was applied to make the fault characteristics more prominent. The specific parameters of the experimental platform are shown in Table 3.

Table 3. Rolling Bearing Parameter Information.

| Parameter | Value |
|---|---|
| Sampling frequency (kHz) | 12.36 |
| Sampling speed (r/min) | 1200, 1400, 1600, 1800 |
| Data acquisition card | National Instruments |
| Bearing diameter (mm) | 15 |
| Rated motor power (kW) | 2.5 |

The preprocessing of vibration signals and parameter settings are identical to those in Section 4.1. With 1024 sampling points as the unit, the time-frequency images generated through continuous wavelet transform have a size of 32×32. Each fault type contains 100 time-frequency image samples. The diagnostic results are shown in Figure 10, and the confusion matrix is shown in Figure 11.

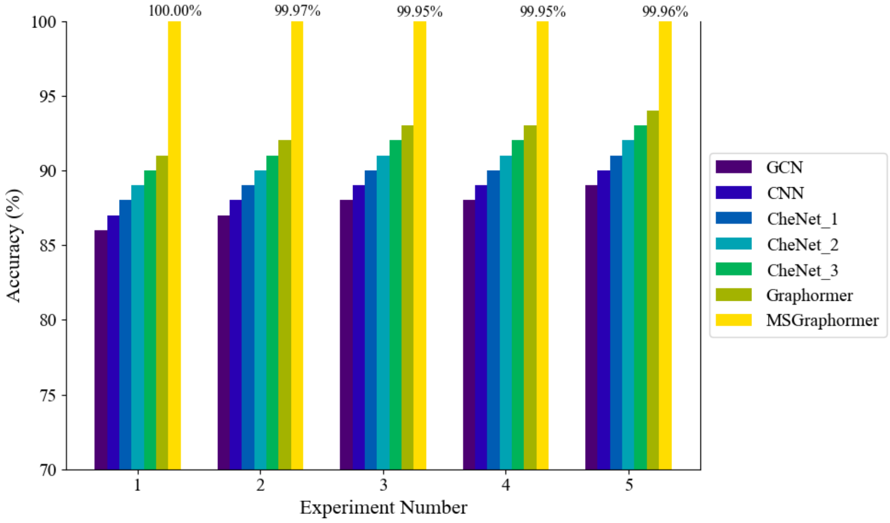


Figure 10. Diagnostic Results of Different Methods for Rolling Bearing Experiments.

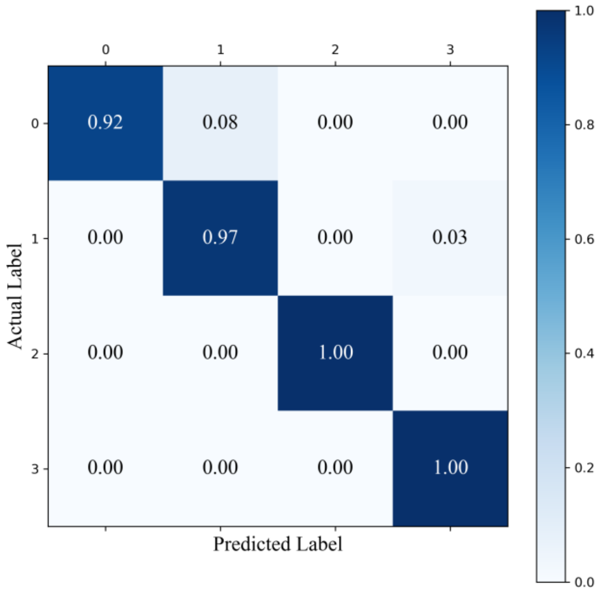


Figure 11. Confusion Matrix of Test Results for Rolling Bearing Experiments.

The highest diagnostic accuracy achieved by our method is 100%, with the lowest being 99.95%, and a standard deviation of 0.03%. This demonstrates the significant advantage of the proposed diagnostic method in terms of both accuracy and stability when compared other methods. Several factors contribute to these results:

(1) Traditional methods like GCN and CNN have simple structures, which limit their ability to extract fault features effectively.

(2) The Chebyshev network shows increased diagnostic accuracy with larger feature extraction scales, indicating that multi-scale feature extraction enhances the model's fault diagnosis capabilities.

(3) The Graphormer leverages the Transformer architecture to focus on critical fault features, thereby improving model accuracy even further.

Our method combines a multi-scale feature extraction module with Transformer-based feature learning, resulting in superior fault recognition rates as compared to other techniques.

1. Noise Resistance Analysis

In practical production, due to severe environmental noise interference, the collected vibration signals are prone to contamination, which requires that the model possess strong noise resistance. Therefore, this paper adds additional Gaussian white noise to the test dataset to verify the noise resistance of the proposed method. Signal-to-noise ratio (SNR) is defined as follows:

$$SNR = 10log_{10}\frac{P_{\text{signal}}}{P_{\text{noise}}} \qquad (12)$$

Where, $P_{\text{signal}}$ and $P_{\text{noise}}$ represent the power of the original signal and the added Gaussian white noise, respectively. In this paper, Gaussian white noise with a signal-to-noise ratio (SNR) ranging from 0 to 9 dB was added to evaluate the proposed fault diagnosis method. The diagnostic results of test samples under different SNRs are shown in Figure 12, where the diagnostic accuracy reached 97.5% even under zero noise interference.

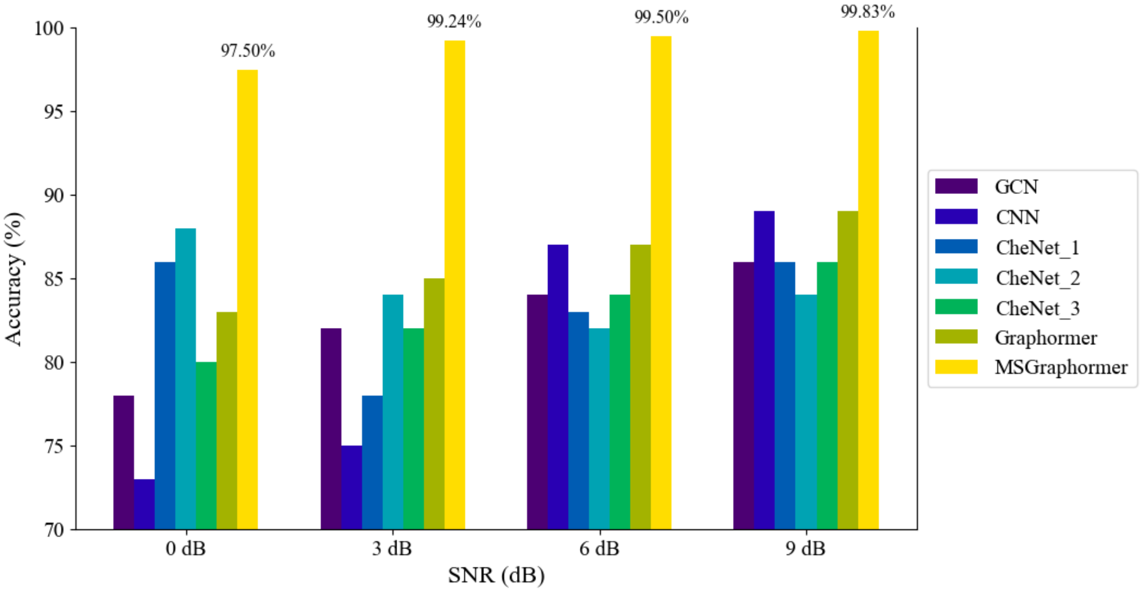


Figure 12. Diagnostic Results under Different Signal-to-Noise Ratios.

As the SNR value increases, the proposed method achieve higher diagnostic accuracy in each test set, with the highest diagnostic accuracy reaching 99.83%. Compared to other methods, the proposed method delivers better diagnostic results under different SNR conditions. This is because the proposed method utilizes multi-scale feature extraction and the Transformer attention module, which enhances the model's ability to aggregate features in noisy environments and adaptively learn the representational relationships between noisy fault samples. Therefore, the noise resistance of the proposed method is notable, maintaining high accuracy under substantial noise interference.

2. Analysis of Generalization Performance

The load of mechanical equipment often varies during actual use. Therefore, to further validate the proposed method's ability to diagnose bearing fault severity under variable operating conditions, this paper selects a load dataset as the training set and three other datasets as the test sets for generalization performance experiments. Specifically, the data under a load of 1200 r/min was used for training, while the data under loads of 1400 r/min, 1600r/min, and 1800 r/min were used for testing. The experimental results are shown in Figure 13, with the highest fault diagnosis accuracy reaching 97.92%, the lowest accuracy remaining at 95.49%, and the average recognition rate being 96.66%. The maximum difference in accuracy is 2.23%. The fault diagnosis accuracy of the proposed method was significantly higher than that of other methods. The reason for this is that the proposed method utilizes multi-scale feature extraction and the Transformer attention module, which enhances the model's ability to aggregate features under variable operating conditions. It effectively extracts internal correlations from fault samples under varying operating conditions. Therefore, the proposed method exhibits strong generalizing capabilities.

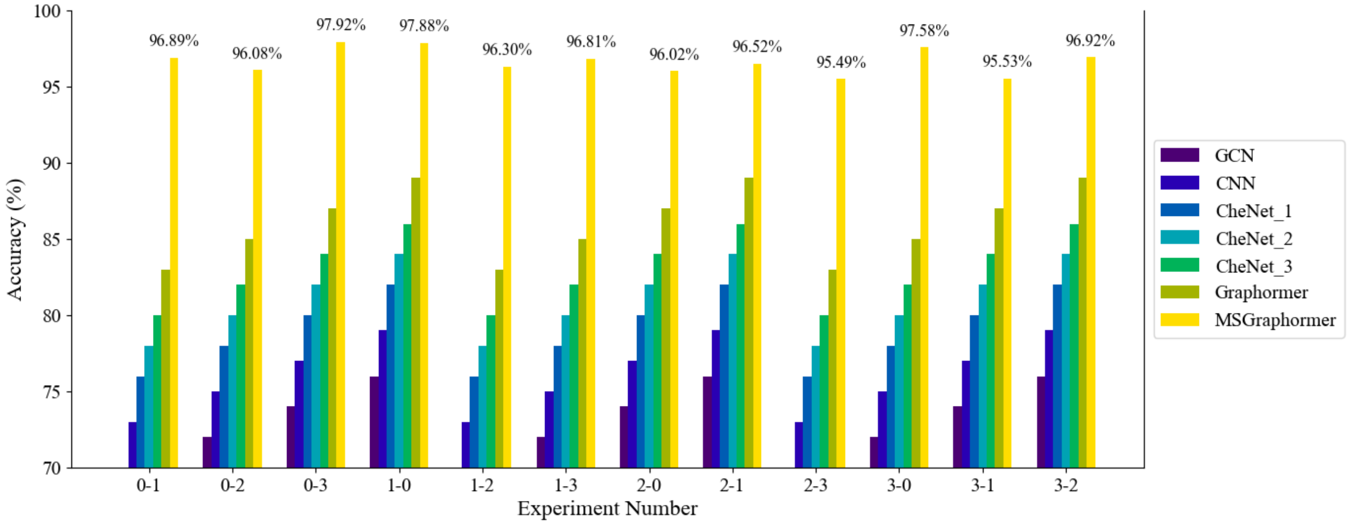


Figure 13. Diagnostic Results under Variable Operating Conditions.

### 4.3. Ablation Experiment

To thoroughly evaluate the performance of the proposed Multi-Scale Graph Transformer (MSGraphormer) in rolling bearing fault diagnosis, we performed a series of ablation experiments. These experiments aimed to analyze the impact of each component of our model and identify the contributions of various features on the overall performance. The key components investigated in our ablation study include the Multi-Scale Feature Aggregation (MSFA) Module, Centrality Encoding, Spatial Encoding, and the Multi-Head Self-Attention Mechanism.

1. Experimental Setup

We utilized the Case Western Reserve University (CWRU) bearing dataset for our ablation experiments, as to the main experiment. The dataset includes vibration signals collected under different fault conditions: normal, inner race fault, outer race fault, and rolling element fault. We performed experiments with the same parameters and preprocessing steps as those described in the main experiment section.

The ablation study was designed to incrementally remove or replace each component of the MSGraphormer model to assess their contribution. The following models were constructed:

(1)Baseline Model (BM): A simple Graph Convolutional Network (GCN) without any of the proposed enhancements.

(2)BM + MSFA: Baseline model with the Multi-Scale Feature Aggregation module.

(3)BM + MSFA + CE: Baseline model with the Multi-Scale Feature Aggregation module and Centrality Encoding.

(4)BM + MSFA + CE + SE: Baseline model with the Multi-Scale Feature Aggregation module, Centrality Encoding, and Spatial Encoding.

(5)Full Model (FM): The complete MSGraphormer with all components.

2. Results and Analysis

The performance of each model variant was evaluated based on its classification accuracy on the test set. The results are summarized in Table 4 and visualized in Figure 14.

Table 4. Ablation Study Results on CWRU Dataset.

| Model | Max Accuracy | Min Accuracy | Average Accuracy ± Std Dev |
|---|---|---|---|
| BM | 82.53% | 78.35% | 80.98% ± 1.64% |
| BM + MSFA | 90.67% | 85.12% | 87.94% ± 2.11% |
| BM + MSFA + CE | 95.33% | 92.47% | 94.12% ± 1.24% |
| BM + MSFA + CE + SE | 97.75% | 95.80% | 96.45% ± 0.98% |
| FM | 99.86% | 99.35% | 99.64% ± 0.10% |

From Table 4 and Figure 14, it is evident that each added component significantly enhances the model's performance. The Baseline Model (BM) achieved an average accuracy of 80.98%, indicating the basic effectiveness of GCNs in fault diagnosis. The addition of the Multi-Scale Feature Aggregation (MSFA) module increased the average accuracy to 87.94%, demonstrating the importance of aggregating information from different neighborhood scales. Introducing Centrality Encoding (CE) further improved the average accuracy to 94.12%, highlighting the value of embedding the structural information of graph nodes. The inclusion of Spatial Encoding (SE) resulted in an average accuracy of 96.45%, indicating the importance of positional information in enhancing feature representation. The

Full Model (FM), incorporating all components, achieved the highest average accuracy of 99.64% with minimal standard deviation, underscoring the robustness and stability of the proposed MSGraphormer. This ablation study clearly demonstrates the significant contributions of each component to the overall performance of the model in rolling bearing fault diagnosis

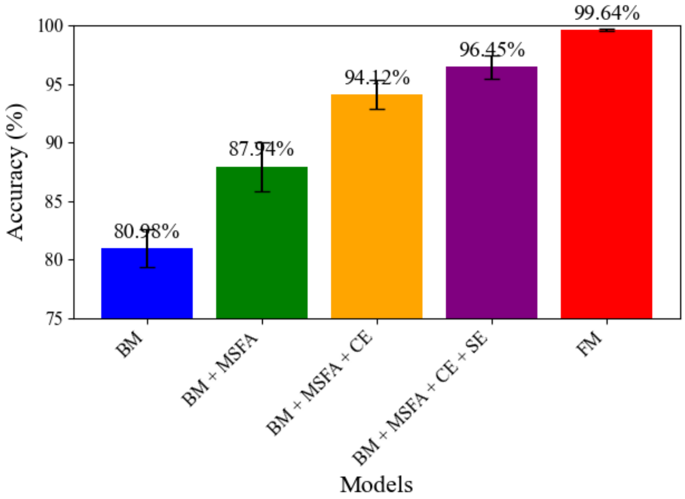


Figure 14. Ablation Study Results Visualization.

## 5. Conclusion

This study proposes a rolling bearing fault diagnosis method based on a multi-scale Graph Transformer and verifies its effectiveness and robustness through extensive experiments. By introducing a multi-scale feature aggregation mechanism, as well as centrality and spatial encoding of graph nodes, this method can deeply explore the structural information of graph nodes and significantly enhance the ability to extract and learn fault features using the multi-head self-attention mechanism.

The experimental results show that the proposed method achieves the highest diagnostic accuracy of 99.5% on the CWRU bearing dataset, with an average accuracy exceeding 97.9%, demonstrating high stability and reliability under different operating conditions. Additionally, on a self-made rolling bearing experimental platform, the method also performs excellently, with the highest diagnostic accuracy reaching 100% and the lowest being 99.95%. These results fully prove the powerful fault diagnosis capability of the multi-scale Graph Transformer method in complex environments.

Through noise resistance analysis, this method exhibits excellent diagnostic performance under different signal-to-noise ratio conditions, with the lowest accuracy being 97.5% and the highest at 99.83%. This indicates that the method has significant advantages in handling noise interference in actual production environments. Meanwhile, generalization performance experiments were conducted to verify the fault diagnosis capability of this method under different loads, achieving the highest diagnostic accuracy of 97.92%, the lowest of 95.49%, and an average recognition rate of 96.66%, demonstrating strong generalization ability. Furthermore, ablation experiments highlighted the significant contributions of each component within the MSGraphormer framework. The integration of multi-scale feature aggregation, centrality encoding, and spatial encoding proved to be beneficial for improving diagnostic accuracy and stability.

Overall, the multi-scale Graph Transformer method proposed in this paper significantly enhances the accuracy, robustness, and adaptability of rolling bearing fault diagnosis. Future research can further optimize feature extraction and computational efficiency to address more practical application scenarios.

## Acknowledgement

This research was supported by the National Science and Technology Major Project (2019-I-0003-0004).

## References

1. Kankar P K, Sharma S C, Harsha S P. Fault diagnosis of ball bearings using machine learning methods[J]. Expert Systems with applications, 2011, 38(3): 1876-1886. DOI: 10.1016/j.eswa.2010.07.119.
2. Lei Y, He Z, Zi Y. EEMD method and WNN for fault diagnosis of locomotive roller bearings[J]. Expert Systems with Applications, 2011, 38(6): 7334-7341. DOI: 10.1016/j.eswa.2010.12.095.
3. Borghesani P, Ricci R, Chatterton S, et al. A new procedure for using envelope analysis for rolling element bearing diagnostics in variable operating conditions[J]. Mechanical systems and signal processing, 2013, 38(1): 23-35. DOI: 10.1016/j.ymssp.2012.09.014.
4. Chen J, Zi Y, He Z, et al. Compound faults detection of rotating machinery using improved adaptive redundant lifting multiwavelet[J]. Mechanical Systems and Signal Processing, 2013, 38(1): 36-54. DOI: 10.1016/j.ymssp.2012.06.025.

5. Jiang H, Li C, Li H. An improved EEMD with multiwavelet packet for rotating machinery multi-fault diagnosis[J]. Mechanical Systems and Signal Processing, 2013, 36(2): 225-239. DOI: 10.1016/j.ymssp.2012.12.010.
6. Pan M C, Tsao W C. Using appropriate IMFs for envelope analysis in multiple fault diagnosis of ball bearings[J]. International Journal of Mechanical Sciences, 2013, 69: 114-124. DOI: 10.1016/j.ijmecsci.2013.01.035.
7. Zhao D, Li J, Cheng W, et al. Compound faults detection of rolling element bearing based on the generalized demodulation algorithm under time-varying rotational speed[J]. Journal of Sound and Vibration, 2016, 378: 109-123. DOI: 10.1016/j.jsv.2016.05.022.
8. Rai A, Upadhyay S H. A review on signal processing techniques utilized in the fault diagnosis of rolling element bearings[J]. Tribology International, 2016, 96: 289-306. DOI: 10.1016/j.triboint.2015.12.037.
9. Cao H, Fan F, Zhou K, et al. Wheel-bearing fault diagnosis of trains using empirical wavelet transform[J]. Measurement, 2016, 82: 439-449. DOI: 10.1016/j.measurement.2016.01.023.
10. Miao Y, Zhao M, Lin J, et al. Application of an improved maximum correlated kurtosis deconvolution method for fault diagnosis of rolling element bearings[J]. Mechanical Systems and Signal Processing, 2017, 92: 173-195. DOI: 10.1016/j.ymssp.2017.01.033.
11. Seera M, Wong M L D, Nandi A K. Classification of ball bearing faults using a hybrid intelligent model[J]. Applied Soft Computing, 2017, 57: 427-435. DOI: 10.1016/j.asoc.2017.04.034.
12. Hoseinzadeh M S, Khadem S E, Sadooghi M S. Modifying the Hilbert-Huang transform using the nonlinear entropy-based features for early fault detection of ball bearings[J]. Applied Acoustics, 2019, 150: 313-324. DOI: 10.1016/j.apacoust.2019.02.011.
13. Hong Y, Kim M, Lee H, et al. Early fault diagnosis and classification of ball bearing using enhanced kurtogram and Gaussian mixture model[J]. IEEE Transactions on Instrumentation and Measurement, 2019, 68(12): 4746-4755. DOI: 10.1109/TIM.2019.2898050
14. Huang R, Li W, Cui L. An intelligent compound fault diagnosis method using one-dimensional deep convolutional neural network with multi-label classifier[C]//2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC). IEEE, 2019: 1-6. DOI: 10.1109/I2MTC.2019.8827030
15. Huang R, Liao Y, Zhang S, et al. Deep decoupling convolutional neural network for intelligent compound fault diagnosis[J]. Ieee Access, 2018, 7: 1848-1858. DOI: 10.1109/ACCESS.2018.2886343
16. Huang R, Li J, Li W, et al. Deep ensemble capsule network for intelligent compound fault diagnosis using multisensory data[J]. IEEE Transactions on Instrumentation and Measurement, 2019, 69(5): 2304-2314. DOI: 10.1109/TIM.2019.2958010
17. Huang R, Wang Z, Li J, et al. A transferable capsule network for decoupling compound fault of machinery[C]//2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC). IEEE, 2020: 1-6. DOI: 10.1109/I2MTC43012.2020.9129078
18. Jin Y, Qin C, Huang Y, et al. Actual bearing compound fault diagnosis based on active learning and decoupling attentional residual network[J]. Measurement, 2021, 173: 108500. DOI: 10.1016/j.measurement.2020.108500.
19. Li J, Huang R, He G, et al. A two-stage transfer adversarial network for intelligent fault diagnosis of rotating machinery with multiple new faults[J]. IEEE/ASME Transactions on Mechatronics, 2020, 26(3): 1591-1601. DOI: 10.1109/TMECH.2020.3025615
20. Huang R, Li J, Liao Y, et al. Deep adversarial capsule network for compound fault diagnosis of machinery toward multidomain generalization task[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 70: 1-11. DOI: 10.1109/TIM.2020.3042300
21. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arxiv preprint arxiv:1409.0473, 2014. DOI: 10.48550/arXiv.1409.0473
22. Song J, Qin X, Lei J, et al. A fault detection method for transmission line components based on synthetic dataset and improved YOLOv5[J]. International Journal of Electrical Power & Energy Systems, 2024, 157: 109852. DOI: 10.1016/j.ijepes.2024.109852.
23. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arxiv preprint arxiv:2010.11929, 2020. DOI: 10.48550/arXiv.2010.11929
24. Zhang Z, Wu L. Graph neural network-based bearing fault diagnosis using Granger causality test[J]. Expert Systems with Applications, 2024, 242: 122827. DOI: 10.1016/j.eswa.2023.122827
25. Zhang J, Cheng Y, He X. Fault Diagnosis of Energy Networks Based on Improved Spatial–Temporal Graph Neural Network With Massive Missing Data[J]. IEEE Transactions on Automation Science and Engineering, 2023. DOI: 10.1109/TASE.2023.3281394
26. Srinivas A, Lin T Y, Parmar N, et al. Bottleneck transformers for visual recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 16519-16529. DOI: 10.1109/CVPR46437.2021.01625

27. Yu J, Li J, Yu Z, et al. Multimodal transformer with multi-view visual representation for image captioning[J]. IEEE transactions on circuits and systems for video technology, 2019, 30(12): 4467-4480. DOI: 10.1109/TCSVT.2019.2947482

28. Chen X, Wang H, Ni B. X-volution: On the unification of convolution and self-attention[J]. arxiv preprint arxiv:2106.02253, 2021. DOI: 10.48550/arXiv.2106.02253

29. Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022. DOI: 10.1109/ICCV48922.2021.00986