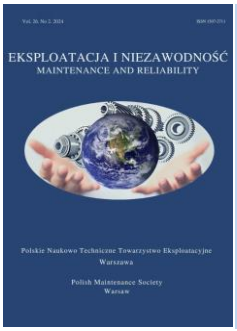


Eksploatacja i Niezawodnosc – Maintenance and Reliability

Volume 26 (2024), Issue 2

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

Article citation info:
Wang X, Yao Y, Gao C, Wasserstein Distance- EEMD Enhanced Multi-Head Graph Attention Network for Rolling Bearing Fault Diagnosis Under Different Working Conditions, Eksploatacja i Niezawodnosc – Maintenance and Reliability 2024: 26(2) http://doi.org/10.17531/ein/184037

Wasserstein Distance- EEMD Enhanced Multi-Head Graph Attention Network for Rolling Bearing Fault Diagnosis Under Different Working Conditions

Indexed by: Web of Science Group

Xingbing Wang[a], Yunfeng Yao[b,*], Chen Gao[c,*]

[a] College of Mechanical and Electrical Engineering, Wenzhou University, China
[b] College of Mechanical and Electrical Engineering, Jiaxing Nanhu University, China
[c] School of Mechanical and Transportation, Jiaxing Nanyang Polytechnic Institute, China

Highlights

- Wasserstein Distance- EEMD is proposed to improve the weights of the node graph.
- Multi-head AE is used in GAT to enhance the stability of the attention-learning process.
- The average classification accuracy of 99.55% is obtained in different working conditions.

Abstract

Traditional fault diagnosis models often overlook the interconnections between segments of vibration data, resulting in the loss of critical feature information. Additionally, the vibration signals of rolling bearings exhibit non-linear behaviors during operation. Therefore, an efficient fault diagnosis model tailored for rolling bearings is proposed in this paper. In the proposed model, the 1D vibration signals are first preprocessed using ensemble empirical mode decomposition (EEMD). This technique generates multiple intrinsic mode functions (IMF) as individual nodes. The percentage distance between each node is calculated using the Wasserstein distance (WD) to capture the relationships between nodes and use it as the edge weights to construct a node graph. This unique approach enhances the transformation of 1D vibration signals into a node graph representation, preserving important information. An improved multi-head graph attention network (MGAT) model is established to extract features and perform classification on the node graph. This MGAT model effectively utilizes the relationships between nodes and enhances the accuracy of fault diagnosis. The experimental results demonstrate that the proposed method achieves higher accuracy compared to similar models while requiring less processing time. The proposed approach contributes significantly to the field of fault diagnosis for rolling bearings and provides a valuable tool for practical applications.

Keywords
ensemble empirical mode decomposition, Wasserstein distance, multi-head graph attention network, fault diagnosis; rolling bearing

1. Introduction

With the rapid development of modern science and technology, enterprises are striving towards high-quality development. In the field of rotating machinery, rolling bearings are widely used and bear important production and safety responsibilities. However, occasional failures in certain parts of rolling bearings caused by harsh environments (such as high temperature, humidity, high speed, and heavy loads) can disrupt the normal operation of mechanical systems and pose economic losses and safety hazards to enterprises 1. Therefore, identifying fault states and accurately diagnosing fault categories in rolling

(*) Corresponding author.
E-mail addresses: X. Wang 21451439035@stu.wzu.edu.cn, Y. Yao yunfeng.yao@jxnhu.edu.cn, C. Gao gaochen_1993@163.com

bearings are of significant importance 2.

In recent years, due to its powerful data processing capabilities, deep learning (DL) has demonstrated remarkable performance in various fields such as natural language processing, computer vision, speech recognition, and healthcare 34. Consequently, deep learning-based approaches for mechanical fault diagnosis have gradually become a hot research topic among domestic and international researchers, yielding many achievements 57. However, traditional deep learning models based on 1D vibration signals tend to overlook the relationship between feature information extracted from each signal segment, limiting the performance of subsequent diagnostic models 8. Zhao et al. 9 introduced a method that converts 1D vibration signals into 2D grayscale images using a sliding window technique. They utilized convolutional neural networks (CNNs) to extract features from the transformed images, enabling effective fault classification. Chen et al. 10 recognized that the sensor signals and their interactions in industrial processes can be represented as graphs with nodes and edges. The authors converted sensor signals into heterogeneous graphs with multiple edge types and employed attention mechanisms to learn adaptive edge weights. The authors employed independent graph neural network blocks to extract fault features from subgraphs of each edge type, followed by a weighted sum function to cascade or fuse the features, resulting in the final graph embedding. Zhou et al. 11 expanded the 1D sensing signal into 2D matrices by minimizing multiscale permutation entropy and encoding it into gray Euclidean distance plots, enriching the information of fault samples. Wang et al. introduced a novel spatio-temporal GNN with an attention-aware module to learn weights flexibly and model the importance and correlation of individual sensors 12. The authors demonstrated the scientific and accurate multi-source information fusion using a wind turbine dataset. Yuan et al. 13 aimed to explicitly capture the features of vibration signals and explore the relationships between signals. They combined the strong feature representation capability of graph attention networks with the recursive nature of non-linear time series to transform vibration signals into recursive graphs. Moreover, the authors proposed a multi-kernel Gaussian symmetric graph attention mechanism to obtain the Hilbert space distribution between neighboring nodes. The effectiveness and superiority of the proposed method under strong noise samples were validated using a wheel-bearing dataset.

However, the above mentioned fault diagnosis models often overlook the interconnections between different segments of the sample 1415. Consequently, critical feature information may be lost. Moreover, it should be noted that the vibration signals of rolling bearings exhibit non-linear behaviors during operation. An efficient fault diagnosis model designed for rolling bearings is designed in this paper to address the aforementioned issues by combining ensemble empirical mode decomposition (EEMD) and a multi-head graph attention network (MGAT). The main work of this paper can be summarized as follows:

(1) A new approach for constructing a node graph using EEMD and WD distance is proposed to improve the binary weighting relationship between nodes by assigning accurate and effective weight values to the edges.

(2) A multi-head attention mechanism is employed in GAT, enhancing the stability of the attention-learning process. The accuracy of fault diagnosis is improved in comparison to traditional methods by effectively utilizing the relationships between nodes.

(3) The effectiveness of the proposed approach is demonstrated through extensive experimentation on the XJ dataset. The average classification accuracy of 99.55% was obtained after five trials, showcasing its robustness across different fault categories under the same working conditions, as well as the same fault category under different working conditions.

2. Theoretical Foundations

2.1. Ensemble Empirical Mode Decomposition

The empirical mode decomposition (EMD) is a data-adaptive, multi-resolution technique used to decompose signals into physically meaningful components proposed by Huang in 1998 16. It is capable of handling non-stationary and non-linear signals. The method recursively decomposes the original signal into components with different resolutions, referred to as intrinsic mode functions (IMF). IMFs are a finite set of functions obtained from the EMD decomposition of the original signal. Each IMF represents a single oscillatory mode that satisfies the following two conditions:

(1) the number of extrema and zero-crossings must be equal

or differ by at most one and

(2) the average value of the envelope formed by the local maxima and minima is zero.

The intrinsic mode functions characterize the inherent oscillatory modes of the signal. The instantaneous frequency of each component can be obtained by performing a Hilbert transform on each IMF component; the frequency component corresponding to any given time point in an IMF is unique. Unlike Fourier transform and wavelet transform, which require pre-defined basis functions, the empirical mode decomposition does not use fixed functions or filters. Based on the characteristics of empirical mode decomposition, this method has two advantages for decomposing and denoising time series data:

1) EMD is suitable for non-linear and non-stationary processes.
2) EMD is adaptive and does not require the selection of predefined basis functions.

In the theoretical framework of EMD, any complex signal can be regarded as a sum of several different and independent IMFs. The decomposition steps of EMD are as follows:

The local maxima points are connected with a cubic spline curve to form the upper envelope, and the local minima points are connected to form the lower envelope, hence identifying all extrema points in the original signal x(t). The upper and lower envelopes include all the data points.

The average values $m_i(t)$ of the upper and lower envelopes are computed:

$$m_i(t) = e_{max}(t) + e_{min}(t)/2. \quad (1)$$

The intermediate signal is computed by subtracting the average envelope curves from the original signal:

$$C_{1,1}(t) = x(t) - m_i(t). \quad (2)$$

It is determined whether the intermediate signal $C_{1,1}(t)$ satisfies the two basic conditions of the IMF. If it satisfies the conditions, it is proven that the signal can be considered an IMF component. Otherwise, steps (1)-(3) are repeated until the conditions are met.

$$C_{1,k-1}(t) - m_i(t) = C_{1,k}(t), \quad (3)$$

$$I_1(t) = C_{1,k}(t), \quad (4)$$

where $I_1(t)$ represents the highest frequency component of the original signal $x(t)$; the remaining component $r_1(t)$ can be obtained by subtracting $I_1(t)$ from the original signal $x(t)$.

The second component $I_2(t)$ can be obtained by continuing the above analysis on the remaining component $r_1(t)$. Subtracting $I_2(t)$ from $r_1(t)$ yields the remaining component $r_2(t)$. This process is repeated iteratively until the last component cannot be further decomposed. The same analysis is then performed on the remaining component $r_1(t)$ to obtain the second component $I_2(t)$. This process of differencing and decomposition continues until the last component cannot be further decomposed.

$$r_1(t) - I_2(t) = r_2(t) \quad (5)$$
$$r_i(t) - I_n(t) = r_n(t)$$

Parameter $r_n(t)$ is considered as the residual component once it becomes a monotonic function, indicating the completion of the entire EMD process. At this point, the sum of all the IMF components and the residual component $r_n(t)$ equals the original signal x(t):

$$x(t) = \sum_{i=1}^{n} I_i(t) + r_n(t). \quad (6)$$

The description of the EMD decomposition process is shown in Figure 1.

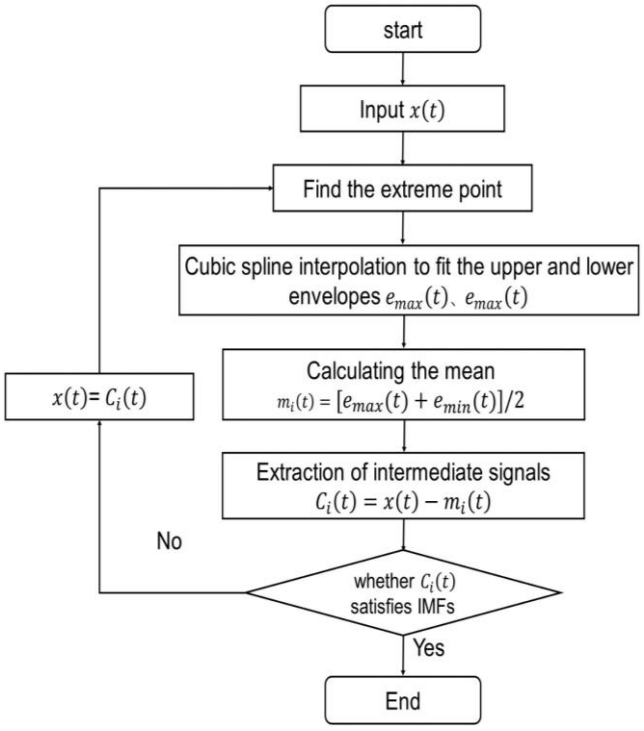


Fig. 1. EMD decomposition process.

However, despite the wide applicability of EMD and its ability to handle signals of various types and trends, it still suffers from two main issues: mode mixing and spurious components 1718. Signals often have strong background noise in real-world working environments, significantly affecting the

quality of EMD decomposition. This, in turn, leads to severe mode mixing problems, where each IMF contains multiple frequency characteristics. Consequently, the IMF is no longer independent but exhibits coupling phenomena. These issues result in the masking of certain features in the signal 19.

Wu et al. introduced the Ensemble Empirical Mode Decomposition (EEMD) method to mitigate the above phenomenon. The EEMD is designed to enhance the decomposition quality by introducing white noise to homogenize the extrema 20. White noise possesses a uniformly distributed spectrum; its addition serves to distribute the signal automatically across appropriate reference scales. The fundamental concept behind the EEMD method involves executing multiple EMDs while superimposing Gaussian white noise. This concept leverages the statistical property of Gaussian white noise, which exhibits a uniformly distributed frequency, to modify the extrema characteristics of the signal by adding distinct white noise instances with equal amplitudes 21. Subsequently, the resulting IMF obtained from multiple EMD iterations is averaged to eliminate the added white noise, suppressing mode mixing. The decomposition procedure of EEMD can be summarized as follows:

1. The total number of iterations is set to *M*.
2. A white noise signal $n_i(t)$ with a standard normal distribution is added to the original signal $x(t)$ to generate a new signal:

$$x_i(t) = x(t) + n_i(t), \tag{7}$$

where $n_i(t)$ represents the white noise sequence added in the *i-th* iteration, and $x_i(t)$ represents the signal with the additional noise in the *i-th* experiment, where $i = 1,2,\dots,M$.

3. EMD decomposition is performed on the obtained noisy signal $x_i(t)$, resulting in the respective IMF and the residual component:

$$x_i(t) = \sum_{j=1}^{J} c_{i,j}(t) + r_{i,j}(t), \tag{8}$$

where $c_{i,j}(t)$ represents the *j-th* IMF obtained from the decomposition of $x_i(t)$ after adding white noise in the *i-th* iteration; $r_{i,j}(t)$ represents the residual function, which represents the average trend of the signal. Lastly, *J* is the number of IMF obtained.

4. Steps (2) and (3) are repeated for M times, each time decomposing the signal with a different amplitude of white noise added and resulting in a set of IMF:

$$c_{1,j}(t), c_{2,j}(t), \dots, c_{M,j}(t), j = 1,2,\dots,J. \tag{9}$$

5. Ensemble averaging is performed on the corresponding IMF mentioned above by utilizing the principle that the statistical average of unrelated sequences is zero, resulting in the final IMF after EEMD decomposition denoted as:

$$c_j(t) = \frac{1}{M}\sum_{i=1}^{M} c_{i,j}(t), \tag{10}$$

where $c_j(t)$ represents the *j-th* IMF obtained from the EEMD decomposition, where $j = 1,2,\dots,J$, $i = 1,2,\dots,M$.

The description of the EEMD decomposition process is shown in Figure 2.

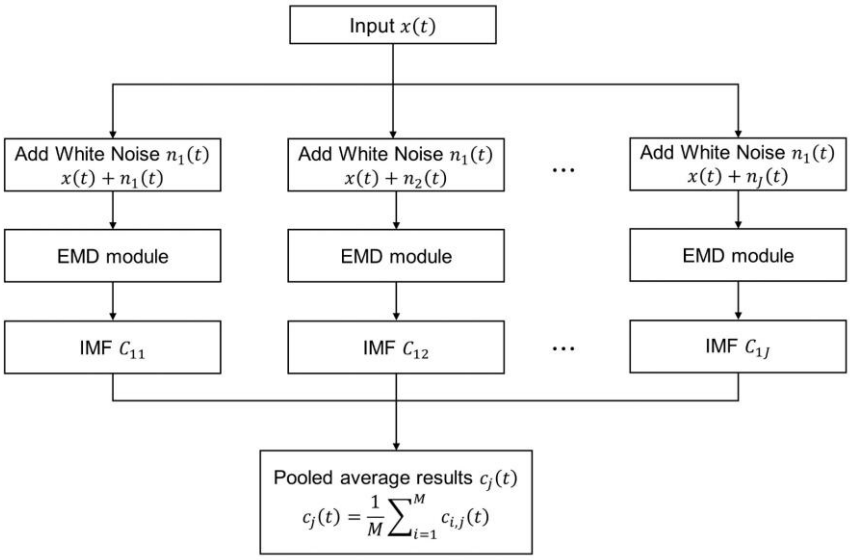


Fig. 2. EEMD decomposition process.

2.2. Wasserstein distance

Wasserstein Distance (WD) is a measure that evaluates the minimum cost required to transform the distribution *P* into the distribution *Q* 22. The value of KL divergence becomes meaningless when the two distributions, *P* and *Q,* are far apart and have no overlap. Consequently, the value of JS divergence becomes a constant, leading to a vanishing gradient during the learning process.

WD represents the similarity between two distributions. Compared to KL divergence and JS divergence, if two distributions have no overlap, WD can still reflect their distance. WD measures the minimum average distance required to move the data from distribution *P* to distribution *Q*, and it is defined as:

$$W(P,Q) = inf_{\gamma \sim \prod(P,Q)} E_{(x,y)\sim y}[\|x-y\|], \quad (11)$$

where $\prod(P,Q)$ represents the set of all possible joint distributions combining P and Q. Parameter $(x,y)\sim y$ can be sampled for each possible joint distribution γ to obtain a pair of samples x and y; the distance between samples $\|x-y\|$ can then be calculated. Therefore, the expected value of the sample pair distance can be computed under the joint distribution γ as $E_{(x,y)\sim y}[\|x-y\|]$. Hence, WD can naturally measure the distance between discrete and continuous distributions. WD provides a measure of distance and a powerful approach to transforming one distribution into another, exhibiting stronger adaptability in representing different distribution states.

2.3. Graph Attention Network

The Graph Attention Network (GAT) is an evolution of the Graph Convolutional Network (GCN). While GCN combines the features of neighboring nodes and the structure of the graph, it limits the generalization ability of the trained model on other graph structures. The key difference between GAT and GCN is that GAT introduces the attention mechanism, assigning larger weights to important nodes 23. GAT is a new neural network architecture based on graph-structured data. Compared to traditional neural networks, GAT considers the relationships between data when processing input data, allowing it to capture and handle the correlations between data more accurately. GAT utilizes a hidden self-attention layer to address the limitations of graph convolutional methods 24.

In GAT, the attention layer can be either a single layer or multiple layers. In this paper, multiple graph attention layers are constructed. The input node features are denoted as $h = \{\overrightarrow{h_1}, \overrightarrow{h_2}, \dots, \overrightarrow{h_N},\ \overrightarrow{h_k} \in R^F\}$, where N represents the number of nodes and F represents the number of features for each node. After passing through the attention layer, a new set of node features $h' = \{\overrightarrow{h_1}', \overrightarrow{h_2}', \dots, \overrightarrow{h_N}',\ \overrightarrow{h_k}' \in R^{F'}\}$ is generated. During this process, the number of features for each node may change. After applying a linear transformation to convert the input features into higher-order features, the attention mechanism is used to assign weights to each node:

$$e_{ij} = \partial(W\vec{h}_i, W\vec{h}_j). \quad (12)$$

In the graph attention layer of GAT, the symbol ∂ represents a shared attention mechanism, e_{ij} represents the attention coefficients indicating the importance of node *j's* features to node i, and W represents the learned weight matrix.

In general, GAT allows each node to participate in the computation of other nodes. Not all nodes in the node graph are first-order neighbors. If only first-order neighbors are considered, it would make the coefficients easily comparable between different nodes. Then, all j options are normalized using the softmax function:

$$\partial_{ij} = softmax_j(e_{ij}) = \frac{exp\,(e_{ij})}{\sum_{k\in N_i} exp\,(e_{ik})}. \quad (13)$$

In GAT, the attention mechanism ∂ is a single-layer feed-forward neural network and is parameterized by a weight vector $\vec{a} \in R^{2F'}$. It uses the LeakyReLU activation function with a negative slope of 0.2 for nonlinearity. When fully expanded, the coefficients computed by the attention mechanism can be represented as follows:

$$\partial_{ij} = \frac{exp\,(LeakyReLU(\vec{a}^T[W\vec{h}_i \| W\vec{h}_j]))}{\sum_{k\in N_i} exp\,(LeakyReLU(\vec{a}^T[W\vec{h}_i \| W\vec{h}_j]))}, \quad (14)$$

where $\|$ represents the concatenation operation, and T represents the transpose operation. The final output feature vector for each node is obtained through a non-linear activation function after calculating the corresponding linear combination of features and obtaining the normalized attention coefficients. The final output feature vector can be represented as follows:

$$\vec{h}_i{}' = \sigma(\textstyle\sum_{j\in N_i} \partial_{ij} W\vec{h}_j). \quad (15)$$

The multi-head attention mechanism is beneficial for making the learning process of attention more stable. The working principle is illustrated in Figure 3. *K*-independent

attention mechanisms are used to concatenate the results output from the equation mentioned earlier, which can be represented as follows:

$$\vec{h_i}' = \|_{k=1}^{K}\sigma(\sum_{j\in N_i}\partial_{ij}^{k} W^{k}\vec{h_j}), \quad (16)$$

where ∂_{ij}^{k} represents the normalized attention coefficients calculated by the k-th attention mechanism and W^{k} is the weight matrix for the corresponding input linear transformation.

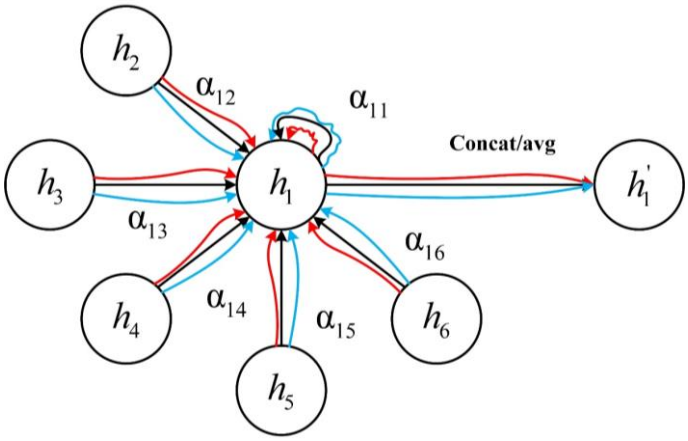


Fig. 3. multi-head graph attention mechanism.

3. The proposed Method

The proposed method, EEMD-WD-GAT, first samples the collected vibration signal by sliding a window to increase the number of samples and avoid the endpoint effect problem in EMD. Then, the collected samples are decomposed into multiple IMFs as nodes by performing EEMD. The feature vectors of each IMF are constructed to represent the information of each IMF. The weights of edges between each node are determined using the WD percentage to measure the relationship between each node, forming structurally complete node graphs for each sample. These node graphs are divided into training, validation, and test sets. Then, a multi-head GAT model is constructed, trained, and validated using the data from the training and validation sets. Finally, the model's performance is tested using the node graphs from the test set.

3.1. EEMD-WD build node graph

The node graph serves as a representation that encapsulates the features of individual nodes and characterizes the relationships among each distinct node, offering a comprehensive depiction of the hierarchical features of the samples and delivering high-quality inputs for subsequent tasks.

In this section, the initially collected 1D vibration signals undergo a sampling procedure using a sliding window to obtain discrete samples. Subsequently, the collected samples are subjected to the EEMD process to derive the Intrinsic Mode Function (IMF) for each individual sample. Next, the feature vectors for each IMF are formulated as node features. The Wasserstein Distance (WD) percentage between each IMF is computed to determine the edge weights. This process defines the constituent nodes and edges of the node graph, as illustrated in Figure 4.

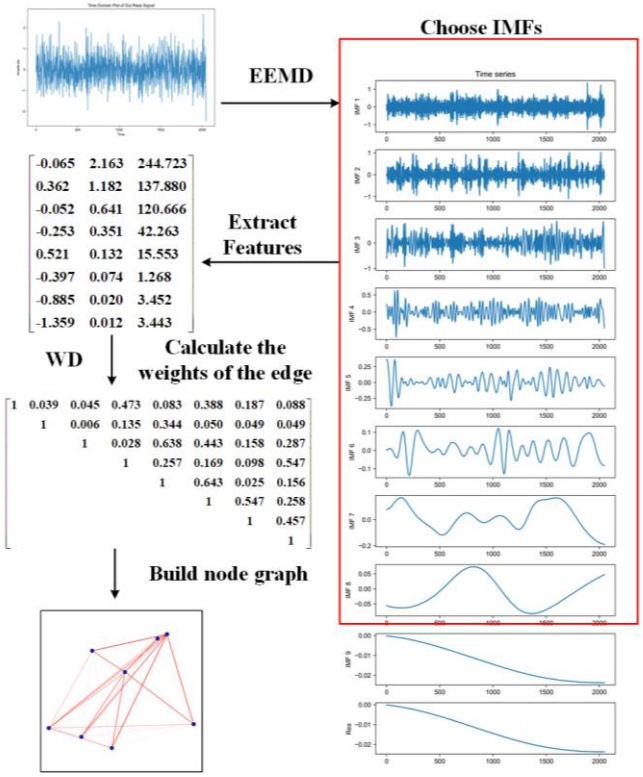


Fig. 4. The process of building a node graph.

The construction of the node graph proceeds as follows:

1. Appropriate sliding window size and sliding step size are selected based on the length and characteristics of the collected 1D vibration raw signal to sample the raw signal. The specific parameters for sliding window sampling are shown in Table 3, and the process is illustrated in Figure 5.

Table 1. Bearing failure classification and failure information.

Original signal length	Windows number	Windows size	step
32768	100	2048	360

2. Performing EEMD on the collected samples yields the IMF components for each sample. The kurtosis, slope ratio at the zero crossing 25, and energy of each IMF are selected to construct the feature vector Q for the node.

$$Q = \{qd, sr, E\}. \quad (17)$$

3. The distance between each IMF component in the sample was calculated using WD. The distance was converted to a percentage share value and then subtracted from that value by one as the weight of the edges in constructing the node graph.

$$g = 1 - \left({wd_{ij}} \Big/ {\sum_1^N wd_i} \right), \tag{18}$$

where wd_{ij} denotes the WD value between nodes i and j, and N denotes the number of nodes in the node graph.

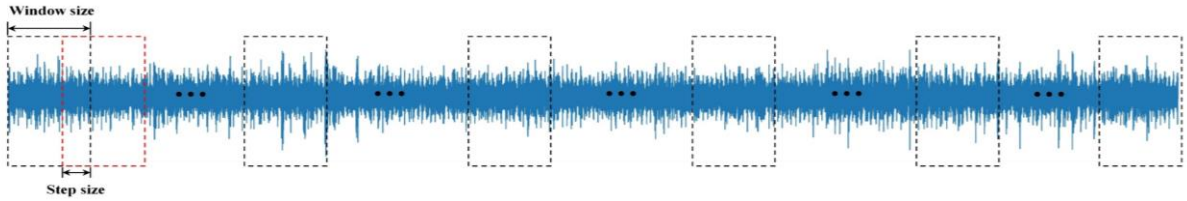


Fig. 5. The schematic diagram of sliding window sampling.

3.2. Fault diagnosis model construction

The basic framework of the GAT network consists of graph filtering layers and non-linear activation layers. Depending on the specific downstream classification task, the output of the graph filtering layer can be used as input for other specific layers. In this paper, the GAT model takes the entire node graph as input to generate new representations; then, it uses these representations to train the classifier for the node graph. The specific process is as follows:

1. The parameters of the GAT model are initialized based on the node graph constructed in the previous steps.
2. The GAT model is pre-trained using the training set samples, and the network model parameters are updated through backpropagation using the output error obtained from the validation set.
3. Softmax classification function is employed to backpropagate the error obtained in Step 2 and update the model parameters using the Adam optimizer. Cross-entropy is selected as the loss function and iterated multiple times until the loss function reaches its minimum value, achieving the best performance of the GAT model. Then, the test set is input into the GAT diagnosis model to complete the fault diagnosis of the rolling bearing.

The rolling bearing fault diagnosis model based on EEMD-WD-GAT is shown in Figure 6.

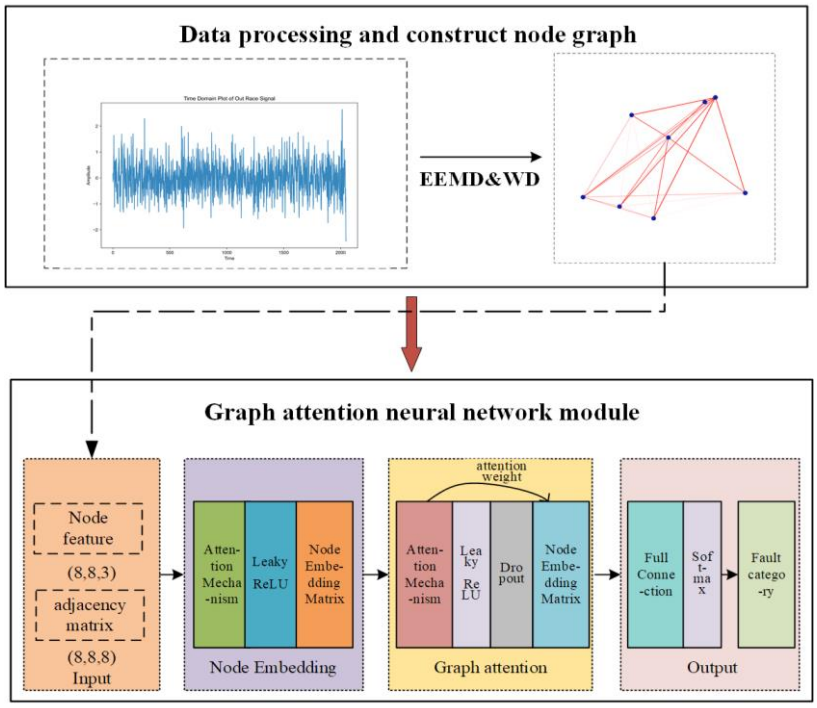


Fig. 6. The schematic diagram of the EEMD-WD-GAT model.

4. Case study

4.1. XJTU-SY dataset

The LDK-UER204 rolling bearing fault dataset, which was acquired from a collaborative laboratory formed by Xi'an Jiaotong University and Shenyang Technology Company, is used in this study. The bearing's parameters are detailed in Table 2. The experimental arrangement comprises an AC motor, a motor speed controller, an accelerometer sensor, a shaft, a hydraulic loading system, and the test bearing, as shown in Figure 7.

Table 2. Bearing failure classification and failure information.

Parameters	Value	Parameters	Value
Inner ring raceway diameter /mm	29.30	Ball number	8
Outer ring raceway diameter/mm	39.80	contact angle /(°)	0
Ball diameter /mm	7.92	Basic dynamic load rating /N	12820

This experimental setup can conduct accelerated life tests for various types of rolling bearings under different operating conditions. The setup allows adjustment of radial load and speed. Acceleration signals in the horizontal and vertical directions of the rolling bearing were collected using an accelerometer sensor. The sampling frequency was 25.6 kHz, the sampling interval was 1 min, and the sampling time per measurement was 1.28 s. Vibration signals of four fault types, including inner race fault, outer race fault, cage fault, and mixed fault, were collected under three different operating conditions. The specific parameters for each fault type are shown in Table 3.

In this study, the vibration signals in the horizontal direction were selected as the raw data. The EEMD-WD-GAT model proposed in this study was used to classify the fault categories under each operating condition and validate its classification capability. In addition, the EEMD-WD-GAT model was applied to handle fault categories across different operating conditions and verify its performance in different condition fault diagnoses 26.

4.2. Parameterization of GAT model

In this paper, the number of attention heads is set to four, the number of iterations of the model is 400, the initial learning rate is 0.0001, and the Adam adaptive optimizer is used as the algorithm to optimize the parameters of the model. The number of samples processed in each batch is eight, and the batch normalization and activation function operations are carried out on each layer of the network. Lastly, the value of dropout is set to 0.6 during training to hide a part of the error weights and prevent the model from overfitting. The node graphs constructed according to EEMD-WD have a feature dimension of 8×3 for each node graph.

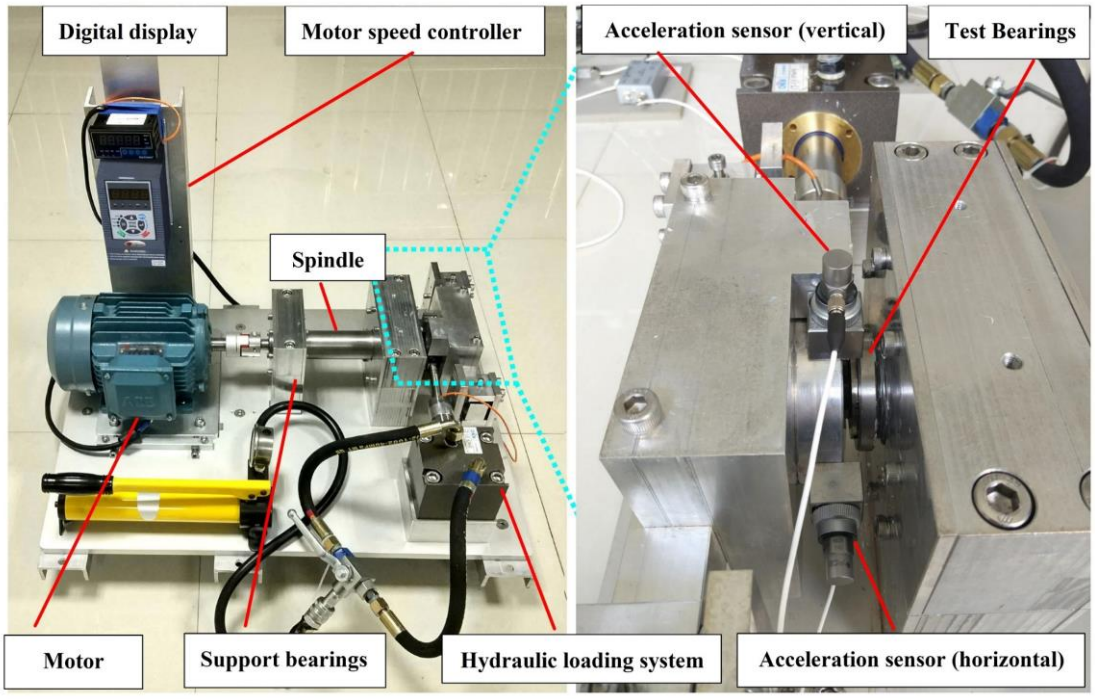


Fig. 7. The bearing test experiment setup of XJTU-SY.

Table 3. Bearing failure classification and information.

	angular velocity	Load	Fault one	Fault two	Fault three
LU-1	2100 rpm	12 kHz	Inner	Cage	Inner and Outer
LU-2	2250 rpm	11 kHz	Inner	Outer	Cage
LU-3	2400 rpm	10 kHz	Outer	Inner	Inner and Cage

4.3. Experimental validation and analysis

4.3.1. Validating the model under the same conditions

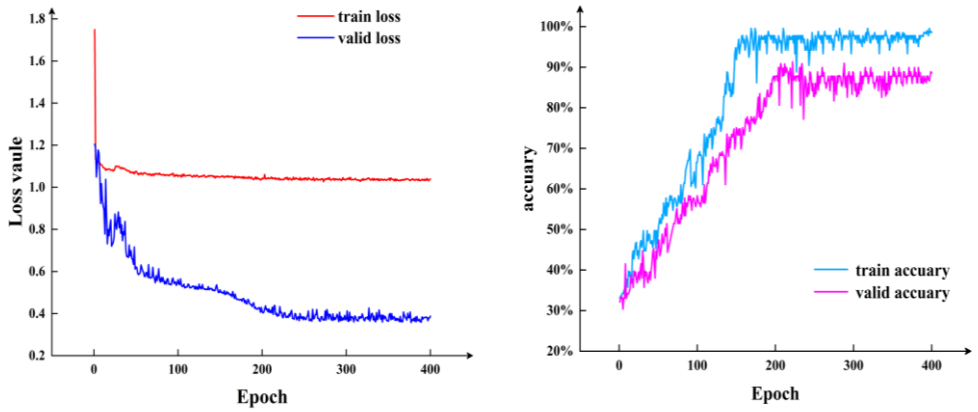
In this part, the three types of faults of LU-3 are analyzed. For each type, 3000 node graphs are selected to divide the training set, validation set, and test set according to the ratio of 8:1:1, labeled as dataset B-0. Then, 100 and 200 inputs from the validation set and the test set are selected into the training set for training, labeled as dataset B-100 and B-200, respectively. The training set is trained according to the model parameters and verified on the validation set. The optimal weight parameters are saved and passed to the test set, and the model is verified on the test set. Simultaneously, this paper also establishes three fault diagnosis models as comparison models to verify the advantages of the model proposed in this paper. The specific parameters of the three types of models are set as follows:

(1) EEMD-SVM: The regularization factor is set to 1.5, and the radial basis kernel function is used.

(2) EEMD-CNN: The EEMD process of the original signal is consistent with the one proposed in this paper. A two-layer CNN model with 16 and 32 convolutional kernels in convolutional layer 1 is set up. The sizes of the convolutional kernels are both 2×1 with a step size of 1, and the sizes of the pooling layer 1 and the pooling layer 2 are 2×1 with a step size of 2. Softmax is used as the activation function, the fully connected layer serves as a classifier, and the model is iterated 200 times.

(3) EEMD-WD-GCN: The node graph is constructed using EEMD-GCN, a two-layer GCN model is designed, and ReLU is used as the activation function for each layer. The output of the first layer is 8×16, and the output of the second layer is 8×32. Adam adaptive optimizer is used as the algorithm to optimize the parameters of the model. The number of samples processed in each batch is 8, softmax is used as the activation function, the fully-connected layer is used as a classifier, and the model is iterated 200 times.

The training and validation loss values of the model under the LU-3 same operating conditions dataset and the accuracies of the training and validation sets are shown in Figure 8.



(a) loss function curve (b) accuracy curve

Fig. 8. The training result of EEMD-WD-GAT model.

According to the analysis of the loss value and accuracy of the model on the training and validation sets, the loss function value of the model on the training and validation sets is stable at approximately 1.04 and 0.42, respectively. The accuracy is stable at roughly 98.28% and 86.34%, and the loss value of the training set of the model is higher than that of the validation set, indicating that the model achieves the expected results.

The accuracy was calculated in the test sets B-0, B-100, and B-200 using the three comparison methods and the proposed model, respectively; the specific results are shown in Table 4

and Figure 9.

According to the analysis of the distribution results of the classification accuracy, EEMD-WD-GAT achieved the best results in different test sets. The model's accuracy was improved by 9.72% over the comparative method (EEMD-SVM) when no samples from the test set were involved in the training. Moreover, there was an improvement of 8.47% in the accuracy of the model when 200 samples were involved in the training, and the classification accuracy reached 99.55%, accomplishing the downstream classification task. The confusion matrix of EEMD-WD-GAT in the B-0, B-100, and B-200 test sets is shown in Figure 10.

Table 4. Accuracy results on the test set under different methods.

Method	Classification accuracy (%)		
	B-0	B-100	B-200
EEMD-SVM	0.8626	0.8948	0.9108
EEMD-CNN	0.8764	0.9025	0.9368
EEMD-WD-GCN	0.9149	0.9522	0.9633
EEMD-WD-GAT	0.9598	0.9866	0.9955

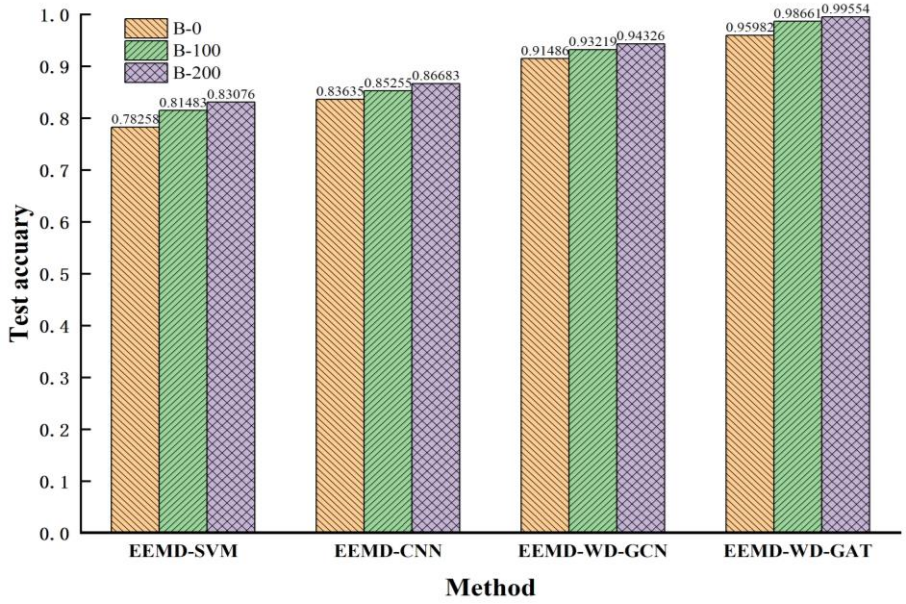
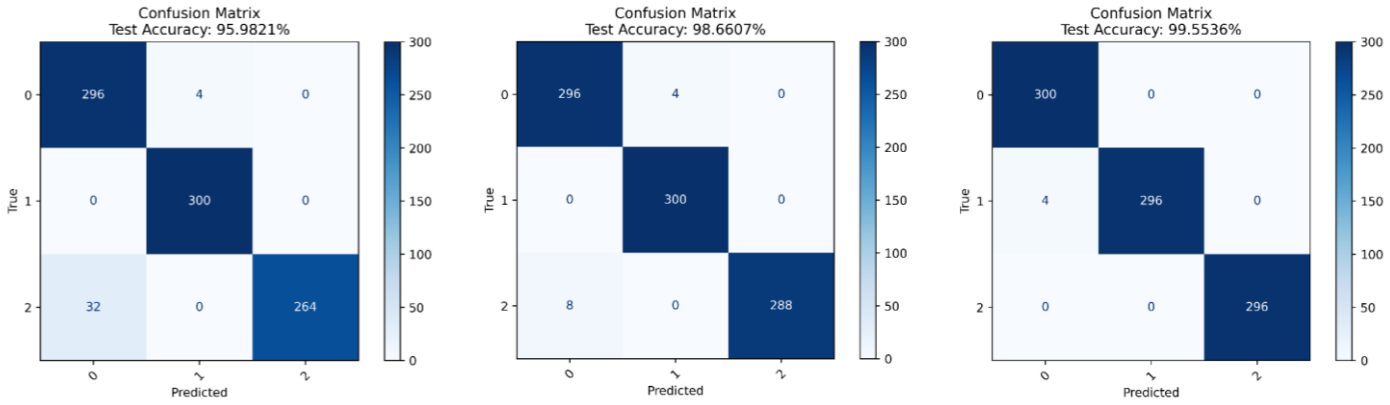


Fig. 9. Accuracy results on the test set under different methods.



(a) B-0 (b) B-100 (c) B-200

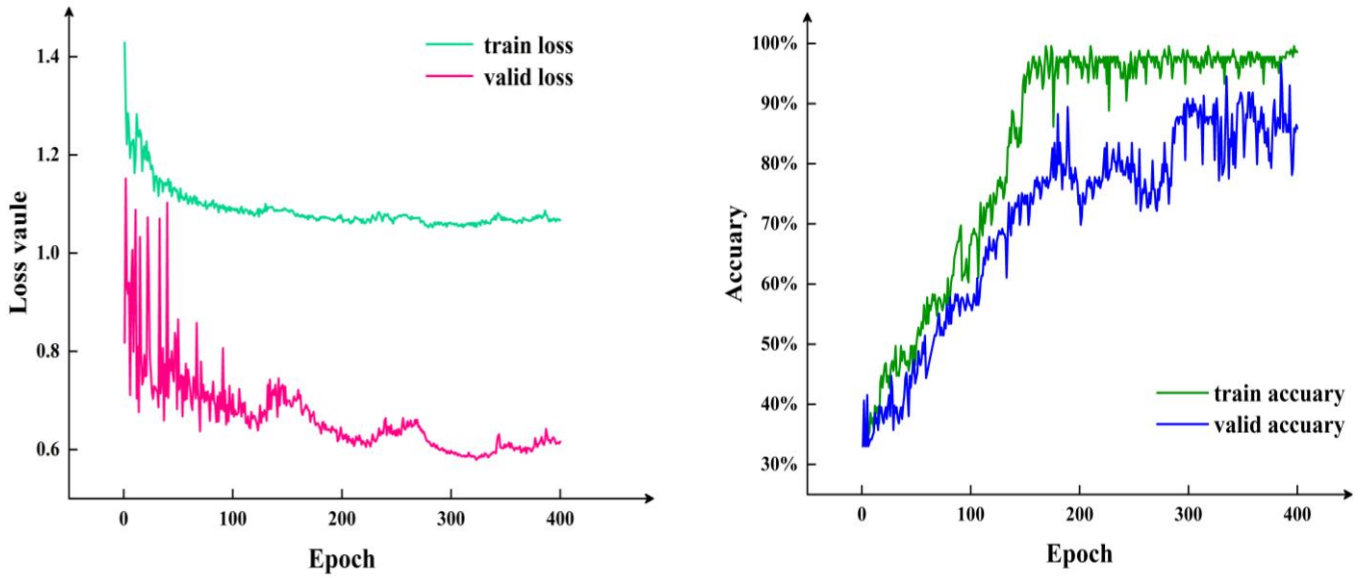
Fig. 10. The confusion matrix of EEMD-WD-GAT.

4.3.2. Validation of model generalization ability under different conditions

In this part, the outer ring faults under LU-1, LU-2, and LU-3 are analyzed under different conditions. Three thousand node graphs are selected as samples for each class, and the training, validation, and test sets are divided according to the ratio column of 8:1:1, labeled as dataset B-0. Then, 100 and 200 inputs from the validation set and the test set, respectively, are selected as the training set for training, labeled as dataset B-100 and B-200, respectively. The training set is trained according to the model parameters and validated on the validation set. The optimal weight parameters are saved to be passed to the test set,

and the generalization ability of the model is verified on the test set. Validation is performed on the three comparative methods proposed in the previous stage and the method proposed in this paper.

According to the analysis of the loss value and accuracy of the model on the training and validation sets, the loss function value of the model on the training set and validation set is stable at about 1.03 and 0.64, respectively. The accuracy is stable at approximately 96.36% and 81.67%, respectively. Moreover, the loss value of the model's training set is higher than that of the validation set, indicating that the model achieves the expected results. The training and validation loss values of the model proposed in this paper under three different working conditions, the accuracy of the training set, and the validation set are shown in Figure 11.



(a) loss function curve (b) accuracy curve

Fig. 11. The training results of EEMD-WD-GAT model different conditions.

The accuracy was calculated using the three comparison methods and the model proposed in this paper in the test sets B-0, B-100, and B-200. The specific results are shown in Table 5 and Figure 12.

Table 5. Classification accuracy on the test set under different methods.

Method	Classification accuracy (%)		
	B-0	B-100	B-200
EEMD-SVM	0.7636	0.7725	0.8135
EEMD-CNN	0.7538	0.7815	0.8026
EEMD-WD-GCN	0.8669	0.8733	0.9347
EEMD-WD-GAT	0.8884	0.9182	0.9955

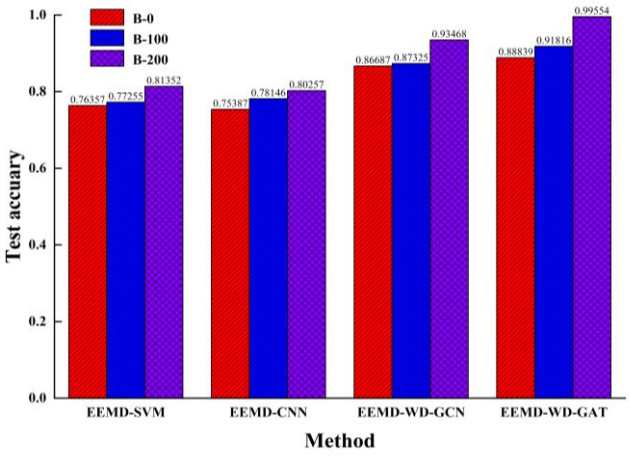


Fig. 12. Accuracy results on the test set B-200 under different method.

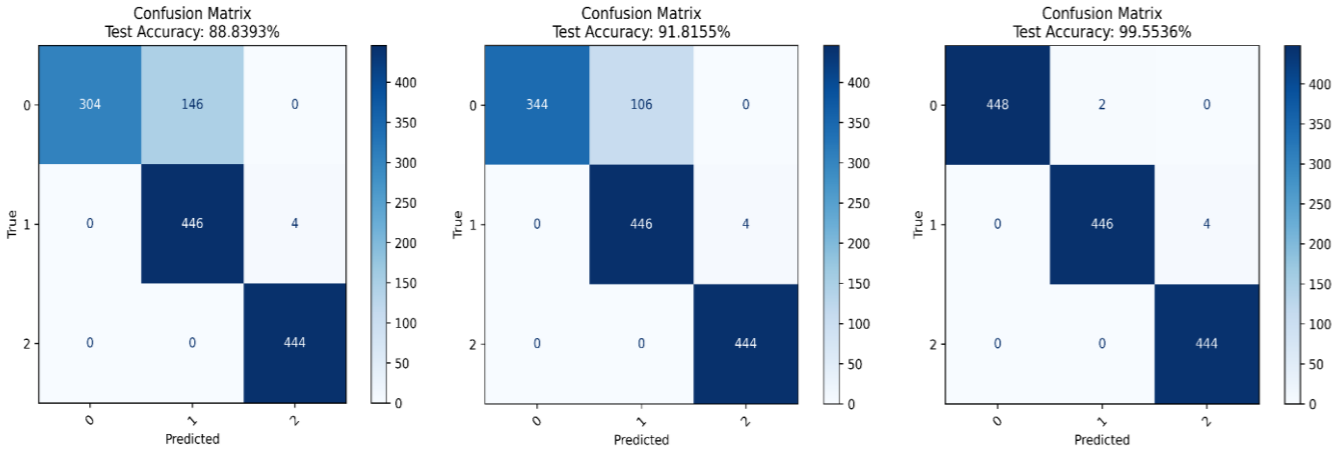
Under heterogeneous conditions, an analysis of classification accuracy distribution reveals that WWMD-WD-GAT consistently outperforms other models across various test sets. Specifically, when no test set samples are included in the training data, WWMD-WD-GAT exhibits a remarkable 12.48%

increase in accuracy compared to EEMD-SVM. Furthermore, even with 200 samples included in the training set, WWMD-WD-GAT demonstrates a substantial 10.71% accuracy improvement, achieving a classification accuracy of 99.55%. This performance addresses the downstream classification task. The confusion matrix for EEMD-WD-GAT on the B-200 test set is shown in Figure 13 to examine the model's performance in detail.

5. Conclusions

(1) This paper proposed a new approach for constructing a node graph using EEMD and WD distance. The WD distance was employed to assess the significance between different nodes, while the correlation between signals was used to determine the weights of the edges, resulting in an adjacency matrix. This new method improves upon the binary weighting relationship between nodes by assigning accurate and effective weight values to the edges. Consequently, the node graph provides high-quality inputs for the GAT model.

(2) This paper conducted experiments on datasets involving different fault categories under the same working conditions, as well as the same fault category under different working conditions, to validate the efficacy and generalization capability of the proposed EEMD-WD-GAT method. The average classification accuracy on the XJ dataset achieved an impressive 99.55% after five trials. This result clearly outperforms the three benchmark methods, confirming the validity and generalization ability of the proposed method. These findings demonstrate the effectiveness and generalization capability of the proposed method, establishing its superiority in terms of accuracy and stability for fault recognition in rolling bearings.



(a) B-0 (b) B-100 (c) B-200

Fig. 13. The confusion matrix of EEMD-WD-GAT in different conditions.

Although this paper has achieved better results in rolling bearing fault diagnosis, there are still the following problems that need to be further studied and solved:

(1) Selecting the effective IMFs after EEMD processing more reasonably and scientifically with additional scientific and rigorous mathematical proof.

(2) designing a more practical loss function for the characteristics of the GAT network model.

In future research, dealing with the connection between node features can be thought of from the following two aspects.

On the one hand, when the correlation, distance, and entropy are discussed from the node features to measure the relationship between features, the research considering the distribution of features will be more scientific and rigorous.

On the other hand, the influence of dimension on the relationship between node features should be considered. For example, after determining the linear relationship between features based on exploring the distribution of features, indicators such as angles should be added to measure the spatial relationship between features. Then, this parameter can be combined with the linear relationship to represent the relationship between node features in space.

Acknowledgement

This research was funded by the Science and Technology Plan Project of Jiaxing in China (Grant. 2021AY10072), and in part by the Research Project of Jiaxing Nanhu University (Grant. 62206ZL) in China.

References

1. J. Lin, H.D. Shao, X.D. Zhou, B.P Cai, B. Liu. Generalized MAML for few-shot cross-domain fault diagnosis of bearing driven by heterogeneous signals. Expert Systems with Applications, 2023, 230, 120696. https://doi.org/10.1016/j.eswa.2023.120696
2. D. Liu, L.L. Cui, W. Cheng. Flexible generalized demodulation for intelligent bearing fault diagnosis under nonstationary conditions. IEEE Transactions on Industrial Informatics, 2023, 19(3): 2717-2728. https://doi.org/10.1109/TII.2022.3192597
3. Y.Q. Zhou, W. Sun, C.Y. Ye, B.H. Peng, X.X. Fang, C. Lin, G.H. Wang, A. Kumar, W.F. Sun. Time-frequency Representation -enhanced Transfer Learning for Tool Condition Monitoring during milling of Inconel 718. Eksploatacja i Niezawodnosc – Maintenance and Reliability 2023: 25(2), 165926. https://doi.org/10.17531/ein/165926
4. D. Liu, L.L. Cui, W. Cheng. A review on deep learning in planetary gearbox health state recognition: Methods, applications, and dataset publication. Measurement Science and Technology, 2023, https://doi.org/10.1088/1361-6501/acf390
5. Z.D. Hei, B.T. Sun, G.H. Wang, Y.J. Lou, Y. Zhou. Multi-feature spatial distribution alignment enhanced domain adaptive method for tool condition monitoring. Eksploatacja i Niezawodnosc-Maintenance and Reliability, 2023: 25(4). https://doi.org/10.17531/ein/171750
6. Y.Q. Zhou, G.F. Zhi, W. Chen, Q.J. Qian, D.D. He, B.T. Sun, W.F. Sun. A New Tool Wear Condition Monitoring Method Based on Deep Learning under Small Samples. Measurement, 2022, 189, 110622. https://doi.org/10.1016/j.measurement.2021.110622
7. S. Yan, H.D. Shao, J. Wang, X.Y. Zheng, B. Liu. LiConvFormer: A lightweight fault diagnosis framework using separable multiscale convolution and broadcast self-attention. Expert Systems with Applications, 2024, 237, 121338. https://doi.org/10.1016/ j.eswa.2023.121338
8. H.C. Wang, W. Sun, W.F. Sun, Y. Ren, Y.Q. Zhou, Q.J. Qian, A. Kumar. A novel tool condition monitoring based on Gramian angular field and comparative learning. International Journal of Hydromechatronics, 2023, 6(2): 93-107. https://doi.org/10.1504/ IJHM.2023.130510
9. J. Zhao, S.P. Yang, Q. Li, Y.Q. Liu, X.H. Gu, W.P. Liu. A new bearing fault diagnosis method based on signal-to-image mapping and convolutional neural network. Measurement. 2021, 176, 109088. https://doi.org/10.1016/j.measurement.2021.109088
10. D Chen, R. Liu, Q. Hu, S.X. Ding. Interaction-Aware Graph Neural Networks for Fault Diagnosis of Complex Industrial Processes. IEEE Trans Neural Netw Learn Syst. 2023, 34(9): 6015-6028. https://doi.org/10.1109/TNNLS.2021.3132376
11. Y.Q. Zhou, H.C. Wang, G.H. Wang, A. Kumar, W.F. Sun, J.W. Xiang. Semi-Supervised Multiscale Permutation Entropy-Enhanced Contrastive Learning for Fault Diagnosis of Rotating Machinery. IEEE Transactions on Instrumentation and Measurement, 2023, 72, 3525610. https://doi.org/10.1109/TIM.2023.3301051
12. Z. Wang, Z.Y. Wu, X.Q. Li, H.D. Shao, T. Han, M. Xie. Attention-aware temporal–spatial graph neural network with multi-sensor information fusion for fault diagnosis. Knowledge-Based Systems. 2023, 10, 110891. https://doi.org/10.1016/j.knosys.2023.110891
13. Z.H Yuan, X. Li, S.Y Liu, Z.Q. Ma. A recursive multi-head graph attention residual network for high-speed train wheelset bearing fault diagnosis. Measurement Science and Technology. 2023, 34(6). https://doi.org/10.1088/1361-6501/acb609
14. Y.Q. Zhou, A. Kumar, C Parkash, G. Vashishtha, H.S. Tang, J.W. Xiang. A novel entropy-based sparsity measure for prognosis of bearing defects and development of a sparsogram to select sensitive filtering band of an axial piston pump. Measurement, 2022, 203, 111997. https://doi.org/10.1016/j.measurement.2022.111997
15. G.X. Zheng, W. Chen, Q.J. Qian, A. Kumar, W.F. Sun, Y.Q. Zhou. TCM in milling processes based on attention mechanism-combined long short-term memory using a sound sensor under different working conditions. International Journal of Hydromechatronics, 2022, 5(3): 243-259. https://doi.org/10.1504/IJHM.2022.125090
16. N.E. Huang, Z. Shen, S.R. Long. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society. A, Mathematical, physical, and engineering sciences. 1998, 454: 903-995. https://doi.org/10.1098/rspa.1998.0193
17. Z.K. Peng, P.W. Tse, F.L. Chu. A comparison study of improved Hilbert–Huang transform and wavelet transform: Application to fault diagnosis for rolling bearing. Mechanical Systems & Signal Processing, 2005, 19(5): 974-988. https://doi.org/10.1016/

j.ymssp.2004.01.006

18. X. Yin, L.Y. Fu. Decorrelation EMD: A new method to eliminate modal aliasing. Vibration and Shock. 2015, 34(4): 25-29. https://doi.org/10.13465/j.cnki.jvs.2015.04.005
19. Y.G. Lei, D.T. Kong, N.P. Li, J. Lin. Adaptive overall average empirical mode decomposition and its application to planetary gearbox fault detection. Journal of Mechanical Engineering. 2014, 50(03). PP 64-70. https://doi.org/10.3901/JME.2014.03.064
20. Z.H. Wu, N.E. Huang. Ensemble empirical mode decomposition: a noise-assisted data analysis method. Advances in Adaptive Data Analysis. 2009, 1(1): 1-41. https://doi.org/10.1142/S1793536909000047
21. H.K. Jiang, C.L. Li, H.X. Li. An improved EEMD with multiwavelet packet for rotating machinery multi-fault diagnosis. Mechanical Systems and Signal Processing. 2013, 36(2): 225-239. https://doi.org/10.1016/j.ymssp.2012.12.010
22. V.M. Panaretos, Y. Zemel. Statistical Aspects of Wasserstein Distances. Annual Review of Statistics and Its Application. 2019, 6(1): 405-431. https://doi.org/10.1146/annurev-statistics-030718-104938
23. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio. Graph attention networks. International Conference on Learning Representations (ICLR) 2018.
24. J. Li, J.R. Wang, H. Lv, Z.X. Zhang, Z.X. Wang. IMCHGAN: Inductive Matrix Completion With Heterogeneous Graph Attention Networks for Drug-Target Interactions Prediction. IEEE/ACM transactions on computational biology and bioinformatics, 2022,19(2): 655-665. https://doi.org/10.1109/TCBB.2021.3088614
25. J.H. Holland. Adaptation in Natural and Artificial Systems:An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence.The MIT Press, Cambridge, Massachusetts, London, England, 1992. https://doi.org/10.7551/ mitpress/1090.001.0001
26. Ya Guo L, Tian Yu H, Biao W, et al. Interpretation of XJTU-SY rolling bearing accelerated life test data set. Journal of Mechanical Engineering. 2019, 55(16): 1-6. https://doi.org/10.3901/JME.2019.16.001