Gao *et al.*
EURASIP Journal on Advances in Signal Processing (2023) 2023:101
https://doi.org/10.1186/s13634-023-01063-6

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

A fault diagnosis method for rolling bearings based on graph neural network with one-shot learning

Yan Gao[1], Haowei Wu[1], Haiqian Liao[1], Xu Chen[2], Shuai Yang[3] and Heng Song[4*]

*Correspondence: songhengyang@163.com

[1] School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China
[2] School of Management Science and Engineering, Chongqing Technology and Business University, Chongqing 400067, China
[3] National Research Base of Intelligent Manufacturing Service, Chongqing Technology and Business University, Chongqing 400067, China
[4] Institute of Management Research, China Railway No.4 Engineering Group, Shanghai 201600, China

Abstract

The manuscript proposes a fault diagnosis method based on graph neural network (GNN) with one-shot learning to effectively diagnose rolling bearings under variable operating conditions. In this proposed method, the convolutional neural network is utilized for feature extraction, reducing loss in the process. Subsequently, GNN applies an adjacency matrix to generate codes for one-shot learning. Experimental verification is conducted using open data from Case Western Reserve University Rolling Bearing Data Center, where four different working conditions with six types of typical faults are selected as input signals. The classification accuracy of the proposed method reaches 98.02%. To further validate its effectiveness, traditional single-learning neural networks such as Siamese, Matching Net, Prototypical Net and (Stacked Auto Encoder) SAE are introduced as comparisons. Simulation results that the proposed method outperforms all chosen methods.

Keywords: Deep learning, Fault diagnosis, Graph neural network, One-shot learning, Rotating machinery

1 Introduction

Rolling bearings, as critical components of wind turbines and various motors, often experience structural damage due to prolonged operation and harsh working conditions. Failure of these bearings can lead to significant losses and severe casualties (Mushtaq et al. 2021). Therefore, it is imperative to detect faults stably and accurately. Due to the complex and harsh working environment, the small fault data of the rolling bearing is difficult to observe directly. Consequently, accurately diagnosing rolling bearing faults has become a prominent research area in equipment fault prediction and maintenance management [13].

Rolling bearing fault diagnosis technology can be broadly categorized into two approaches: signal processing-based diagnosis technology and artificial intelligence-aided pattern recognition-based diagnosis technology. Signal processing-based methods rely on domain expertise to extract fault components from initial noise signals using techniques such as short-time Fourier transform (STFT), wavelet packet

transform (WPT) [31], empirical wavelet transform (EWT) [38], Hilbert–Huang transform (HHT) [35], etc.

Training a substantial deep learning network from scratch requires a huge amount of labeled samples and time. Large-scale data can be trained to develop a classifier that can predict the data distribution, which is also called the large sample learning method. However, in the real-world operation environment, rolling bearing usually works in the normal state. Consequently, obtaining a substantial amount of labeled fault data is challenging [28]. By employing a meta-learning based few-shot learning approach, relational clustering is generated and nearest neighbor classification. Subsequently, classification predictions are produced, enabling the construction of a model that captures class differences limited number of data samples. As such, this paper utilizes the aforementioned method to address the challenge of working with small datasets and achieve higher accuracy in classification.

There are various types of meta-learning, such as neural network initialization parameters, feature spaces suitable for measuring data distances, neural network structures and network parameters, hyperparameters of models, optimizers of neural networks, etc. Few-shot learning is an application of meta-learning in the field of supervised learning. Algorithms based on meta-learning can be classified as the following: metric-based learning methods, model-based methods and optimization-based methods. Among them, metric-based learning methods perform small-sample classification tasks by directly comparing feature metrics through a certain feature space. The GNN model in this paper is based on the metric learning approach to perform the task of small sample classification.

This paper applies the GNN with a one-shot learning method to the fault diagnosis of rolling bearings of rotating machinery devices in an innovative way. CWRU bearing data set is used as the experimental data set, and Siamese network [10, 32], matching network [3], prototype network [2], SAE [20] and other traditional one-shot learning neural network models are selected for comparison. It is verified that the method has higher accuracy of fault diagnosis and stronger generalization ability, making its classification accuracy as high as 98.02%.

Below is a summary of the essential contributions of our work:

1. Utilizing a meta-learning based few-shot learning method to address the limited data in real working conditions and achieve higher accuracy in classification.
2. Employing a GNN model based on a metric learning approach for few-shot classification tasks.
3. Innovatively applying the G a one-shot to fault diagnosis of rotating machinery devices' rolling bearings.

The remaining sections of this paper are structured as follows. Section 2 presents an overview of related literature. Section 3 introduces the theoretical basis and the whole process of the proposed method. Section 4 evaluates the proposed approach through mechanical equipment fault diagnosis experiments. Section 5 provides results with a discussion of the case. Conclusions and future works for the work are drawn in Sect. 6.

2 literature review

Compared to traditional fault diagnosis methods based on signal processing and analysis, an increasing number of researchers have applied deep learning to fault diagnosis, resulting in remarkable achievements. The utilization of deep learning partially reduces the reliance on prior knowledge, simplifies the signal processing procedure and enhances the accuracy of fault diagnosis.

Standard deep learning neural network models include deep belief networks (DBNs) [19], convolutional neural networks (CNNs) [1], SAE [14] and recurrent neural networks (RNNs) [27]. Shao et al. [25] used double-tree complex wavelet packet (DTCWPT) to extract the fault characteristics of the original vibration signal. They designed an adaptive depth confidence network for rolling bearing fault diagnosis. Good results have been achieved in this way. [17] introduced CNN into bearing fault diagnosis and carried out comparative tests, which proved that the proposed method is superior to the traditional support vector machine (SVM) that directly inputs signals into the classifier. This method has higher diagnostic accuracy. Ince et al. [9] applied one-dimensional convolution neural network (1DCNN) to the early fault diagnosis of the motor. [33] proposed an end-to-end rolling bearing fault diagnosis model based on 1DCNN, which can achieve high-precision diagnosis in a noisy environment without using denoising pretreatment. [36] applied the combination of 1DCNN and antagonism adaptive in the cross-domain fault diagnosis of rolling bearings. [18] used the generic adversary network (GAN) to generate a small number of fault samples for unbalanced samples, input the synthesized samples into the fault diagnosis model of the stacked denoising auto encoder (SDAE), and achieve good results. Wang et al. [34] combined the denoising auto encoder (DAE) and GAN into gear fault diagnosis. The model has good anti-noise capability and good diagnostic performance. Shao et al. [23] used an automatic encoder to compress data and constructed a bearing fault diagnosis model based on a convolutional deep belief network (CDBN). [5] proposed a rolling bearing fault diagnosis method based on short-time Fourier transform and CNN. Experiments have verified that this method has high recognition accuracy for different types of faults. The above methods often need enough labeled samples to ensure the accuracy of fault diagnosis classification. Highly accurate diagnostic results are often not accurately obtained when faced with fault samples with small amounts of data.

In recent years, GNN [30] has been effective in processing data with rich relational structures but less data volume. GNN was first proposed by [22]. Its purpose is to establish a neural network based on graph theory for the data stored in the graph domain. GNN aggregates the characteristics of neighboring nodes through information transmission between neighboring nodes, which can effectively represent the complex relationship between data. Therefore, the graph field can provide more information than the general data field. The application of GNN to few-shot learning meets the requirements of processing structural information between data in the process of one-shot learning.

Compared with the traditional CNN method, GNN has advantages in processing the identification feature extraction of signals in the discrete space domain. In recent years, GNN method has been successfully applied in many research fields, such as website recommendation system [21], protein molecular structure and performance design [11]. In

the field of fault diagnosis, [12] successfully applied graph neural networks to fault diagnosis of industrial process networks.

3 The proposed method

3.1 Using STFT for data preprocessing

In the research of rolling bearings, time–frequency imaging technology is employed as a foundation for data feature extraction. In this paper, STFT is used to preprocess the vibration signal due to its effectiveness in analyzing time-varying and non-stationary signals [24]. By applying a time-limited window function before the Fourier transform of the signal, it is assumed that the non-stationary signal is stable within each short-time analysis interval. Through sliding the window function along the time, the signal is analyzed segment by segment to obtain a set of local 'spectra' representing its spectrum at different moments in time. After obtaining the spectrum information of the signal in the time domain, filtering can be performed on the signal. The primary frequency components can be directly spectral information, while secondary frequency components considered as noise are eliminated through inverse transformation.

Unstable signals can be processed in the following ways. The center position of the window function is at $t = \mathcal{I}\_0$, The signal is windowed:

$$y(t) = x(t) * r(t - \mathcal{I}_0) \tag{1}$$

t is the time period,$x(t)$ is a small segment of the transformed signal. $r(t - \mathcal{I}_0)$ is the sequence of fundamental window functions. The intercepted signal $y(t)$ can be obtained by multiplying the window function and the original signal. The intercepted signal $y(t)$ is the signal of execution time corresponding to t.

By the Fourier transform, Eq. (1) can be rewritten as [24]:

$$\mathrm{STFT}_f(t,\omega) = \mathcal{F}(x(t)) = \int_{\infty}^{+\infty} f(t) * \quad r(t - \mathcal{I}_0)e^{-\mathrm{jwt}}\mathrm{d}t \tag{2}$$

$\mathrm{STFT}_f(t,\omega)$ is the spectral distribution of the first segmented sequence. From the above formula, it can be seen that with the change of t, all segments will complete the transformation and be combined into a complete signal transformation.

For the convenience of expression, the function $S(\omega,\mathcal{I})$ is defined as:

$$S(\omega,\mathcal{I}) = \mathcal{F}(x(t) * r(t - \beth)) = \int_{\infty}^{+\infty} x(t) * \quad r(t - \mathcal{I}_0)e^{-\mathrm{jwt}}\mathrm{d}t \tag{3}$$

The function $S(\omega,\mathcal{I})$ represents the spectral result $\mathrm{STFT}_f(t,\omega)$ after the transformation of the original function when the window function center is $\mathcal{I}\_0$.

The spectral energy relationship of time can be determined by [24]:

$$\delta_{\mathrm{SP}}(\omega,\ \mathcal{I}) = \|S(\omega,\ \beth)\|^2 = \left\| \int_{\infty}^{+\infty} x(t)^* r(t - \beth_0)e^{-\mathrm{jwt}}\mathrm{d}t^2 \right\| \tag{4}$$

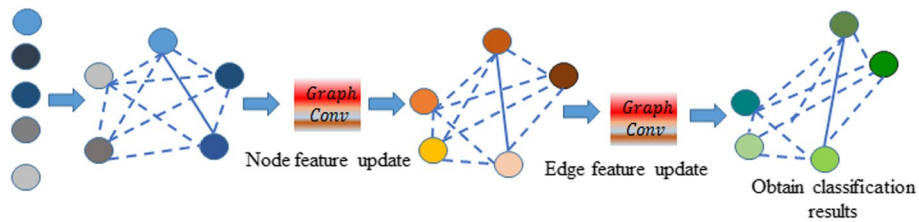

Fig. 1 The typical structure of the GNN

The one-dimensional data can be converted into the picture form of the two-dimensional data by the above method. All raw data were converted into image data as shown in Fig. 1. Image data were divided into two parts. Output data were formed by stacking information such as image, label, one-hot corresponding to the label and its class in the dataset.

3.2 Using CNN to extract feature vectors

Since its inception by [15], CNN has significant advancements in terms of network depth and the development of relevant theories and structures. The extraction process of CNN is mainly to convolve or pool the matrix as a grayscale, RGB image or one-dimensional time-series vibration signal.

In CNN, nodes in the same layer are independent of each other, but neural nodes in different layers are connected in the form of weight sharing. The weights here are obtained through training, that is, W and b, which determine the mapping relationship. When the training data are input into CNN, the convolution layer first performs a convolution operation. The mathematical expression of the convolution process is [4]:

$$z_i^l = f\left(\sum_{j=1}^{J} W_{(i,j)}^l \otimes x_j^{l-1} + b_i^l\right) \tag{5}$$

where z_i^l depicts the lth feature diagram of the lth convolution layer. x_j^{l-1} depicts the jth feature diagram of layer $l-1$. $W_{(i,j)}^l$ corresponds to the jth weight matrix of the ith feature diagram of the ith layer; b_i^l denotes the bias vector. $f()$ denotes the activation function. The conversion of the $l-1$th layer feature diagram to the ith layer is achieved by the convolution process.

The primary activation function used in this paper is the ReLU function. Let the data to be activated be x. Then, its mathematical expression is as follows:

$$\mathrm{ReLU}(x) = \begin{cases} 0, & \text{if } x < 0, \\ x, & \text{if } x \geq 0. \end{cases} \tag{6}$$

Standard pooling methods include maximum pooling and random pooling, which are used to reduce the data dimensions. This paper mainly uses the maximum pooling method, whose mathematical expression is as follows:

$$S_i = \mathrm{sub}_{\mathrm{sampling}}(S_{i-1}) \tag{7}$$

After the data are processed by multiple convolution layers and pooling layers, high-dimensional feature data will be obtained, which needs to be input into the complete connection layer for flattening and processing into feature vectors.

The training and testing process of all samples in this paper is implemented on the Python platform based on Python language. Make the data format of the sample build into $32 \times 32 \times 3$. It is used for the input layer of the neural network. Complete four convolutions, three layers of pooling, and finally output $1 \times 1 \times 64$ feature vector. Then, combine the feature vector with the vector representing the label (one hot code) as the input node of GNN, and input it into GNN together.

3.3 One-shot learning of GNN for fault diagnosis

The GNN model is built upon connection relationships, enabling the extraction of exceptional graph features from both graph nodes and their interconnections [26]. GNN uses vertex updating to reduce the differences between sample features of the same category and increase the differences between sample features of different categories; edge update is used to calculate the similarity between vertices. That is, the attention weight of the features of neighbor vertices is aggregated when the vertices are updated. Therefore, GNN is applied to one-shot learning to meet the requirements of structural processing information between data in the one-shot learning process.

In this paper, each node represents an image. The weight of each edge represents the relationship between the two images (distance or similarity) [16, 29]. Specific weight calculation process [6]:

$$\tilde{A}_{\mathrm{i},j}^{(k)} = \varphi_{\tilde{\theta}}\left(x_i^{(k)}, x_j^{(k)}\right) = \mathrm{MLP}_{\tilde{\theta}}\left(\mathrm{abs}\left(x_i^{(k)} - x_j^{(k)}\right)\right) \tag{8}$$

where x^k is received as input of a GNN layer for graph convolution, where $\varphi = \mathrm{MLP}$ denotes the similarity metric module implemented using a multilayer perceptron structure. In this paper, a multilayer perceptron stacked is considered after the absolute difference between two vector nodes.

After the adjacency matrix is obtained by the above method, the following layer network can be calculated by GNN to complete the GNN transfer. The calculation process is as follows:

$$x_l^{(k+1)} = Gc\left(x^{(k)}\right) = \rho\left(\sum_{B\varepsilon A} Bx^{(k)}\theta_{B,l}^{(k)}\right), l = d_1 \ldots d_{k+1}, \tag{9}$$

where $x_l^{(k+1)}$ represents the characteristics of layer $k+1$ vertex l; Gc represents the graph convolution operation, which can be calculated according to the k layer vertex $x^{(k)}$ to calculate the $k+1$ layer vertex $x^{(k+1)}$. A is the adjacency matrix; B is the relation matrix participating in vertex updating; ρ is a vertex update function with arguments $\theta_{B,l}^{(k)}$. The update rule for the node feature can be calculated using this formula.

$$\Theta = \left\{\theta_1^{(k)}, \ldots, \theta_{|A|}^{(k)}\right\}_k, \theta_A^{(k)} \in \mathbb{R}^{d_k * d_{k+1}} \tag{10}$$

represents the set of training parameters. The accumulation symbol indicates that the adjacency matrix B can adopt a variety of calculation methods and add them together. According to this formula, the update rule of the node feature can be obtained.

As shown in Fig. 1, due to the denseness of the edges in the graph, depth is simply interpreted as giving the model more expressive power. Vertex update refers to sample feature update, which updates vertices according to similarity and category, so as to improve their category generalization ability and prediction accuracy [7]. The essence of edge updating is to calculate the similarity matrix between samples. This matrix is also an attention-weight matrix, which can be used for subsequent vertex aggregation processing. It is updated by means of similarity measurement. In the training process, it is also necessary to change the weight of each network layer. The input $v^{(k)}$ and the output of the graph convolutional block are cascaded to generate the lower-level network input $v^{(k+1)}$.

The feature of the initial point is defined as:

$$x_i^{(0)} = (\phi(x_i), h(l_i)) \tag{11}$$

where $\phi(\cdot)$ is a convolutional neural network, $h(\cdot)$ represents the translation of the tag into a one-hot vector.

The final loss function is [6]:

$$\min \frac{1}{L} \sum_{i \le L} \ell(\Phi(\daleth_i; \theta), Y_i) + \mathcal{R}(\theta) \tag{12}$$

where $\daleth_i$ is the i-th time of the task, the targets Y_i are associated with image categories of designated images $x_i, \ldots x_j \in \daleth_i$ with no observed label. $\{(\daleth_i, Y_i)_i\}_{i \le L}$ is a training set.

$$(\Phi(\daleth; \theta), Y) = -\sum y_k \log P(Y_* = y_k | \daleth)$$

where $\Phi(\daleth; \theta) = P(Y|\daleth)$, the predicted label is obtained through maximum likelihood estimation.

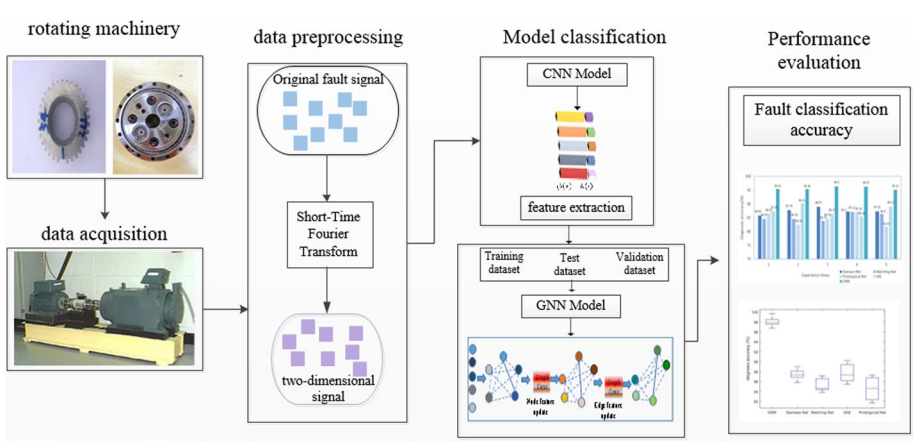

Fig. 2 The working flowchart of the proposed method

The combination of GNN and one-shot learning in meta-learning can be applied to fault diagnosis of rotating machinery. The details are illustrated in Fig. 2 and summarized below.

The methodology employed in this study is listed as follows: firstly, collect signals from the public dataset of CWRU under health and failure conditions, reflecting the actual situation from multiple dimensions. Secondly, the fault state signal of one-dimensional rotating machinery is transformed into a two-dimensional picture by a STFT. Thirdly, feature vectors are extracted by using CNN. Fourthly, these feature vectors are fed into the graph neural network for learning with only a few samples available. Finally, fault diagnosis and classification are obtained.

4 Experiments

4.1 Datasets

The data set from the Case Western Reserve University (CWRU) Rolling Bearing Data Center was used for the bearing fault data with the experimental platform as shown in Fig. 3. The CWRU dataset was obtained using accelerometers to collect vibration data from test-rig consisting of a torque transducer, electronic control equipment, 2-HP motor and dynamometer. The test platform tests the bearings that support the motor. The bearing fault status is measured by EDM technology. (1) Normal Baseline Data; (2) 12 k Fan End Bearing Fault Data;12 k Drive End Bearing Fault Data; (3) 48 k Drive End Bearing Fault Data; (4) 12 k Drive End Bearing Fault Data. In addition, since ORF is a stationary fault, different fault placements located at 3 o'clock (ORF-3), 6 o'clock (ORF-6) and 12 o'clock (ORF-12) were also considered. Four fault diameters (i.e., 0.007 inches, 0.014 inches, 0.021 inches and 0.028 inches) were pre-planted for IRF, RBF, ORF-3, ORF-6 and ORF-12, separately. The data file (mat format) provided by CWRU is limited in length, so overlapping sampling is adopted to generate more time-domain signal samples. We randomly selected 40 classes to form the training dataset. The remaining classes were randomly divided into a validation dataset (10 classes) and a test dataset (10 classes). Each type of motor bearing data collected 300 images for a total of 300*60 pictures.

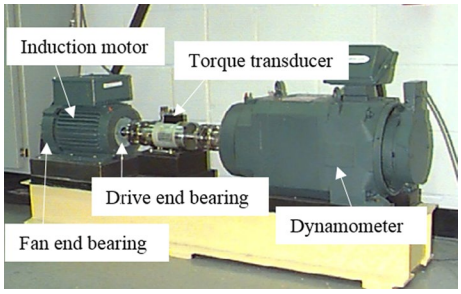

Fig. 3 The experimental setup of bearing (CWRU)

Table 1 Specific types of rolling bearing faults used in the experiment

Damage location	Workload			
	0hp/1797 rpm	1hp/1772 rpm	2hp/1750 rpm	3hp/1730 rpm
Nothing	Nor_0	Nor_1	Nor_2	Nor_3
Inner Race	IR007_0	IR007_1	IR007_2	IR007_3
	IR014_0	IR014_1	IR014_2	IR014_3
	IR021_0	IR021_1	IR021_2	IR021_3
Ball	B007_0	B007_1	B007_2	B007_3
	B014_0	B014_1	B014_2	B014_3
	B021_0	B021_1	B021_2	B021_3
Outer Race Centered@6:00	OR-6-007_0	OR-6-007_1	OR-6-007_2	OR-6-007_3
	OR-6-014_0	OR-6-014_1	OR-6-014_2	OR-6-014_3
	OR-6-021_0	OR-6-021_1	OR-6-021_2	OR-6-021_3
Outer Race Orthogonal @3:00	OR-3-007_0	OR-3-007_1	OR-3-007_2	OR-3-007_3
	OR-3-014_0	OR-3-014_1	OR-3-014_2	OR-3-014_3
	OR-3-021_0	OR-3-021_1	OR-3-021_2	OR-3-021_3
Outer Race Opposite @12:00	OR-3-007_0	OR-3-007_1	OR-3-007_2	OR-3-007_3
	OR-3-021_0	OR-3-021_1	OR-3-021_2	OR-3-021_3

Table 1 shows the specific types of rolling bearing failures used in the experiment, such as IR007_0, IR indicates that the damage position of the fault is the inner race, and 007 indicates that the damage degree of this type of fault is 7 mils, _0 means its workload is 0 hp.

4.2 Experimental environment

The hardware and software settings of this experimental environment are shown in Table 2.

4.3 Model design

During the entire training process, the optimizer selects the Adam optimizer, and the loss function selects the cross-entropy loss function. The batch size is 100. Batch normalization is used in all hidden layers to accelerate training [8], with a fixed learning rate of 0.001. The retention probability settings for Dropout are 0.4 and 0.5, respectively. The graph neural networks in this article use LeakyReLU.

In this paper, the CNN model uses a 5-layer convolutional neural network, including 4 layers of convolution (Conv2d) and 1 layer of linearity. Among them, except for the second layer of convolution without padding, the other three layers are all Conv2d

Table 2 Development environment

Hardware	Configuration or installation information
Processor	Intel(R) Core(TM) i7-6700 CPU @3.40GHZ 3.40 GHz
RAM	64.0 GB
Operating system	Windows 7 Professional 64-bit
Compilation environment	MATLAB R2018a

convolutions with a convolutional kernel size of 3×3, a stride of 1 and a padding of 1. The input of this network is a $(3 \times 84 \times 84)$ image, and the output is a feature vector of length 128.

The GNN model in this paper uses three graph convolution layers, with each graph convolution layer having five convolutions. Among them, each convolutional layer has a convolutional kernel size of 1×1, a convolutional stride of 1 and no padding. The network has an input of $133 \times 6 \times 6$ and an output of a feature vector of length 5. The reddest goes through a LogSoftMax classifier to obtain the individual class probabilities.

5 Results and discussion

In practice, the load of rotating machinery often changes. Therefore, it is necessary to verify the classification accuracy of the proposed fault diagnosis model under different load conditions. A total of 12,000 samples of CWRU datasets were collected from the experiment. The image preprocessed by STFT is shown in Fig. 4.

As shown in Fig. 5, the loss function initially exhibits a relatively large value, reaching 1.78 as observed in the above figure, while achieving an accuracy of only 3% for the graph neural network model. However, after undergoing 350 iterations, there is in fault classification accuracy and a particularly prominent reduction in the loss function. It is noteworthy that both fault classification accuracy and loss function values stabilize

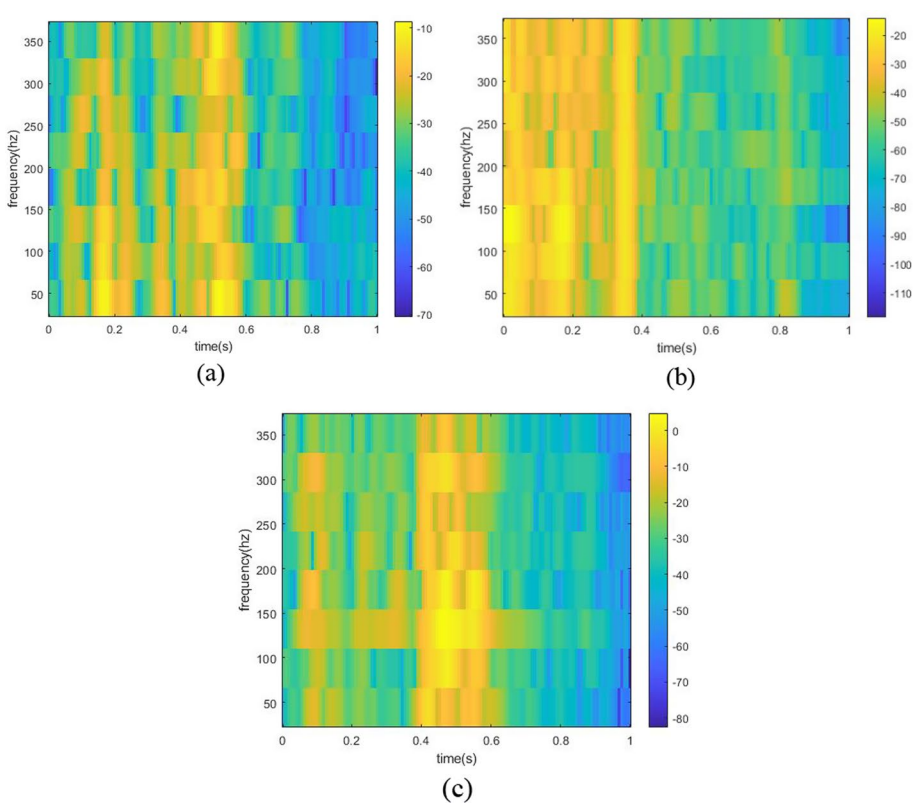

Fig. 4 Signals acquired from bearings of CWRU: **a** Fan end bearing fault data of inner race and **b** Fan end bearing fault data of ball; **c** Drive end bearing fault data of ball(The paper uses 60 types of failure, due to space limitations, only three are shown here)

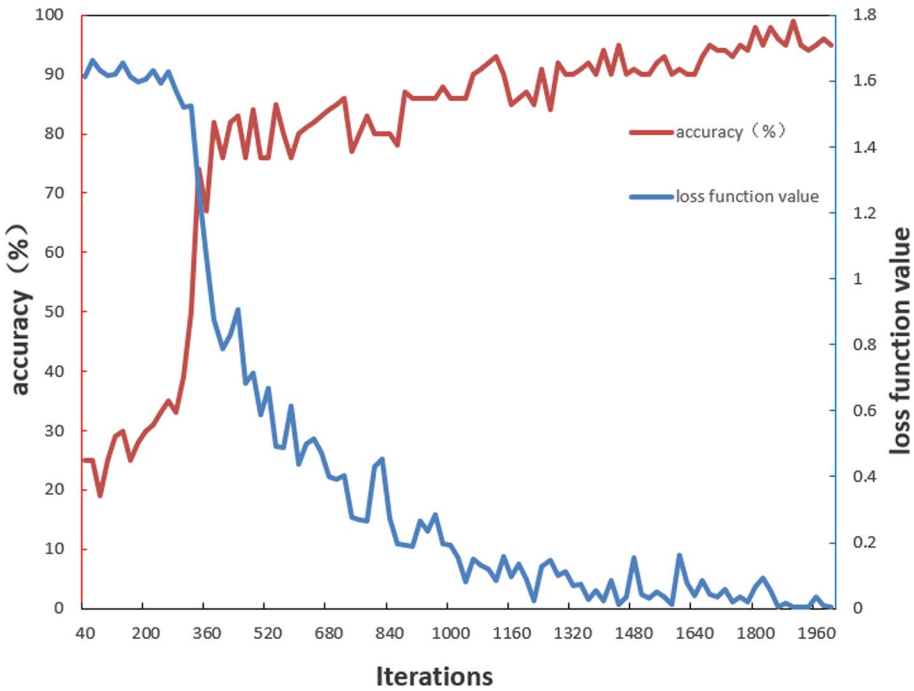

Fig. 5 Classification accuracy of few-shot learning on CWRU

around 1480 iterations. By the time we reach 2000 iterations, the accuracy reaches an impressive 98.02%, with the loss value mere 0.005 effectiveness of the GNN-based fault diagnosis model employed in this experiment.

In the experiment, N-way K-shot was used as the evaluation index. The model was evaluated by performing 1-shot, 5-way experiments on the dataset.

5-way 1-shot, 2-shot, 3-shot, 4-shot and 5-shot were performed on CWRU datasets in Fig. 6. The accuracy of 5-way 5-shot (Accuracy of 99.25%) is higher than that of 5-way 1-shot (Accuracy of 98.02%).

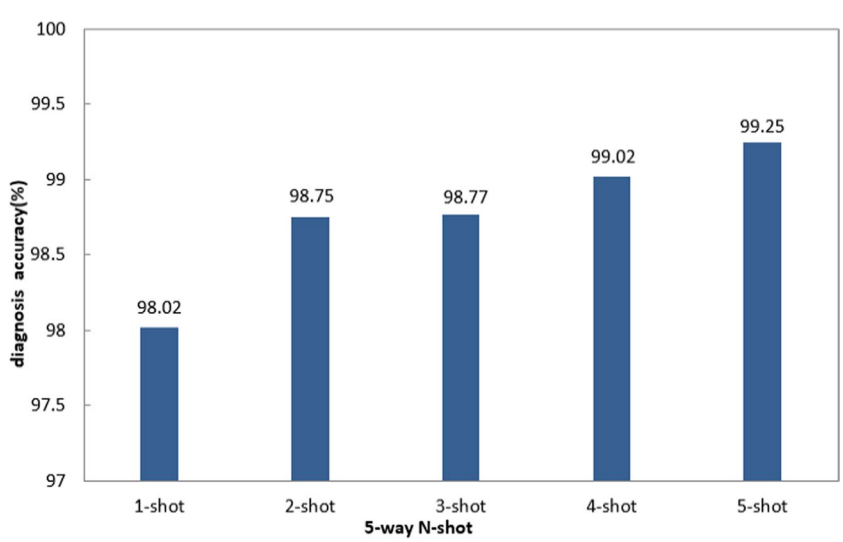

Fig. 6 Classification accuracy of few-shot learning on CWRU

For few-shot fault diagnosis problems, the number of shots (i.e., the value of K) diametrically reflects the intricacy of the problem structure. The smaller the value of K, the more challenging the learning task becomes in terms of accuracy [37]. The accuracy of 1-shot and 5-shot classification tasks has been significantly improved. Since the model can learn more representative features from more labeled samples for classification. This also proves that optimizing correlation between graph nodes and samples enhances measurement label prediction of unknown sample tags, excellent classification accuracy can be achieved.

Among the various few-shot learning methods, the metric learning method in meta-learning stands out as a significant direction. It is characterized by its simplicity and effectiveness, eliminating the need for complex recursive networks and reducing memory requirements. For the new small sample data, the model solves the target problem in a completely feedforward manner without updating the model, which is more convenient for low-latency or low-power application scenarios. The similarity metric-based embedding learning method uses data from embedding layers such as CNN, and the acquired data features can be used as the basis for similarity metric calculation to achieve the classification task. Since metrics such as Euclidean distance and cosine similarity can be used for similarity metrics, the training and learning of this method focuses on feature extraction based on embedding, and can also be extended to the field of metric learning. Siamese Net, Matching Net and Prototypical Net in few-shot learning are the embodiment of this idea. This paper analyzes and compares with representative metric-based few-shot learning models (matching networks, relational networks, and Siamese networks), showing the advantages of graph neural networks in metric-based and meta-learning methods, and also solving the problem of not being able to model data in non-Euclidean spaces, leading to new research directions in the field of few-shot learning classification.

At present, the above comparison models have been widely used in the fields of pattern recognition such as human image recognition, fingerprint recognition and target tracking. Siamese network extracts features through two networks with the same structure. Classification of the described input images is performed based on similarity. Then, the neural network is learned using various types of number of samples and loss functions. After learning the model, the model can make predictions for new samples. Siamese neural network is applied in a simple way for few-shot learning, which is not suitable for unsupervised learning environments. However, the network inspires subsequent-related metric-based models. Matching network introduces an attention mechanism and external memory and uses the nearest neighbor method with end-to-end vectors to classify the samples to be identified and obtain the corresponding classification results. Matching network can complete fast learning and improve the generalization ability of the trained samples and enhance few-shot learning performance. Matching networks cannot solve the problem of memory being over-occupied. For this problem, a new modeling method, prototype network, is proposed. Prototype network uses convolutional neural networks to fuse different types of data into a data fusion approach. The memory consumption is greatly reduced and the accuracy of classification is improved. The straightforward and efficient design of the prototype network model has been widely

Table 3 Results of several methods' categorization using the CWRU dataset

Model	Accuracy of each test (%)					Mean (%)
	1	2	3	4	5	
Siamese Net	85.82	87.79	88.97	87.20	87.33	87.42
Matching Net	84.57	84.59	83.79	87.12	86.40	85.29
Prototypical Net	86.53	82.54	84.59	87.25	81.67	84.52
SAE	87.32	90.23	86.37	85.46	89.12	87.70
GNN	96.73	97.82	99.63	98.12	97.79	98.02

used in many small sample task fields. Encoders are prone to overfitting, and the restriction of sparse expressions is introduced in the self-coding machine to constitute a sparse autoencoder with strong generalization performance. SAE can obtain more abstract and typical compression characteristics from raw data, which improves the performance of traditional self-encoders and shows more practical application value. Therefore, Siamese Net, Matching Net, Prototypical Net and SAE were used for comparison.

Table 3 shows the comparison of the different models. Obviously, GNN had the highest classification rate each time (98.02% on average). The accuracy of SAE is 87.70%. The accuracy of Siamese Net is 87.42%. The accuracy of Matching Net is 85.29%. The accuracy of Prototypical Net is 84.52%. It indicates that GNN is the best classification method for rolling bearing fault diagnosis.

For better validation, a line chart is applied for comparisons. It is a statistical chart that can directly reflect the difference in accuracy. Different models are selected for model diagnosis performance comparison experiments, and the accuracy of fault diagnosis is shown in Fig. 7. GNN algorithm has the highest fault diagnosis accuracy, followed by SAE distribution. GNN is better than that of Matching Net, prototypical Net, Siamese Net and SAE in the density or fault accuracy of the four.

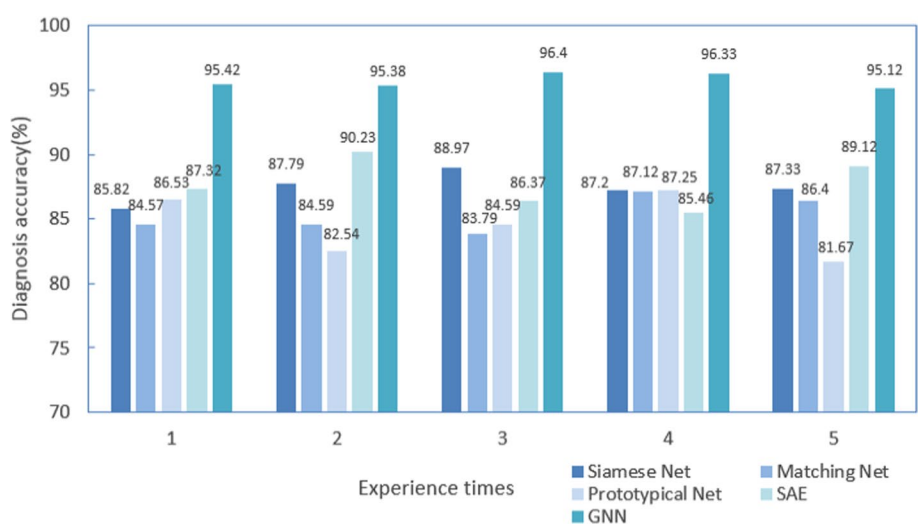

Fig. 7 Classification accuracy of few-shot learning on CWRU

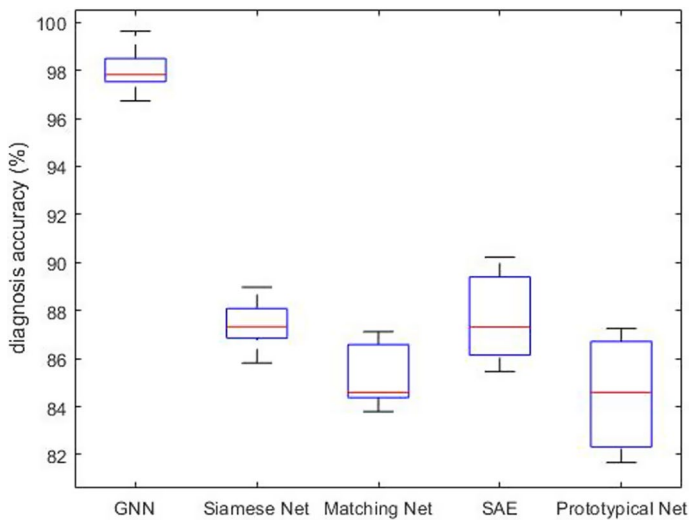

Fig. 8 Box-plot of fault diagnosis results by different algorithms of CWRU dataset

The box-plot can also express the diagnostic performance and stability of the model under different comparison models more intuitively, as shown in Fig. 8, the box plot position of the method in this paper is the highest, and the box plot is the flattest. It shows that this method has the best classification effect and stability compared with other models.

6 Conclusions and future works

To solve the problem of shortage of training sets for rotating machinery faults in practical applications, a one-shot learning for rotating machinery diagnosis method based on graph neural network (GNN) is proposed in this paper. The local frequency spectrum in a small period near time t is obtained by acquiring the rolling bearing signals and performing the STFT on them. The feature extraction is achieved by the CNN. Subsequently, the GNN aggregates under the guidance of node features and edge features, updates edge features through similarity calculation, and finally produces classification results from edge features. Experiments were carried out to verify the efficiency of the proposed method. Siamese Net, Matching Net, prototypical Net and SAE are chosen as the comparisons. The results indicate the proposed method outperforms all the selected methods. The overall accuracy of the proposed method can reach 98.02%.

The proposed method is only chosen to verify the rotating machinery in this paper. Further exploration is still needed to determine the fault diagnosis efficiency of this method for other mechanical devices with small data samples. Therefore, transfer learning will be applied to the proposed method in our future research, allowing the direct application of the GNN fault diagnosis model for rotating machinery to other mechanical devices.

Funding
The Chinese National Natural Science Foundation (51905058), the Chongqing Municipal Education Commission's Science and Technology Research Program (KJZD-K202100804), the Venture & Innovation Support Program for Chongqing Overseas Returnees (cx2021075), and the project of open competition mechanism to select the best candidates of China Railway Group Limited (2021-major-14).

Availability of data and materials
All data generated or analyzed during this study are included in this published article [and its supplementary information files].

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 29 July 2023 Accepted: 2 October 2023
Published online: 11 October 2023

References

1. M. Ali, J.H. Shah, M.A. Khan, M. Alhaisoni, U. Tariq, Brain tumor detection and classification using pso and convolutional neural network. Comput. Mater. Contin. **73**(3), 4501–4518 (2022)
2. M. Bilardo, G. Fraisse, M. Pailha, E. Fabrizio, Design and experimental analysis of an integral collector storage (ICS) prototype for DHW production. Appl. Energy **259**, 114104 (2020). https://doi.org/10.1016/j.apenergy
3. Y. Ding, X. Tian, L. Yin, X. Chen, S. Liu, Multi-scale relation network for few-shot learning based on meta-learning, in *International Conference on Computer Vision Systems* (Springer, Cham, 2019), pp. 343–352
4. W. Fuan, J. Hongkai, S. Haidong et al., An adaptive deep convolutional neural network for rolling bearing fault diagnosis. Meas. Sci. Technol. **28**(9), 95005 (2017)
5. H.Z. Gao, L. Lin, X.G. Chen, Feature extraction and recognition for rolling element bearing faultutilizing short-time fourier transform and non-negative matrix factorization. Chin. J. Mech. Eng. **28**, 96–105 (2015)
6. V. Garcia, J. Bruna. Few-shot learning with graph neural networks (2017). https://doi.org/10.48550/arXiv.1711.04043
7. A. Ghasempour, M. Martinez-Ramon, Electric load forecasting using multiple output gaussian processes and multiple kernel learning, in *IEEE Symposium on Industrial Electronics & Applications (IEEE ISIEA)* (2022)
8. A. Ghasempour, M. Martinez-Ramon, Short-term electric load prediction in smart grid using multi-output Gaussian processes regression, in *IEEE Kansas Power and Energy Conference (IEEE KPEC)* (2023)
9. T. Iince, S. Kiranyaz, L. Eren, Real time motor fault detection by 1-D convolutional neural networks. IEEE Trans. Ind. Electron. **63**(11), 7067–7075 (2016)
10. S. Javed, M. Danelljan, F.S. Khan, M.H. Khan, M. Felsberg, Visual object tracking with discriminative filters and siamese networks. A Survey and Outlook, CoRR http://arxiv.org/abs/2112.02838, pp. 1–20 (2021)
11. S. Kearnes, K. Mccloskey, M. Berndl, Molecular graph convolutions:moving beyond fingerprintsJ. J. Comput. Aided Mol. Des. **30**(8), 595–608 (2016)
12. H. Khorasgani, A. Hasanzadeh, A. Farahat, Fault detection and isolation in industrial networks using graph convolutional neural networks C, in *2019 IEEE International Conference on Prognostics and Health Management (ICPHM)*, California (2019) pp. 1–7
13. P. Kumar, R. Tiwari, Development of a novel approach for quantitative estimation of rotor unbalance and misalignment in a rotor system levitated by active magnetic bearings. Iran J. Sci. Technol. Trans. Mech. Eng. **45**, 769–786 (2021). https://doi.org/10.1007/s40997-020-00364-7
14. S. Karthic, S. Manoj Kumar, Wireless intrusion detection based on optimized lstm with stacked auto encoder network. Intell. Autom. Soft Comput. **34**(1), 439–453 (2022)
15. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
16. Y. Li, D. Tarlow, M. Brockschmidt, R. Zemel, Gated graph sequence neural network. Proc. Int. Conf. Learn. Representat. **23**(1), 63–83 (2015)
17. C. Lu, Z. Wang, B. Zhou, Intelligent fault diagnosis of rolling bearing using hierarchical Convolutional network based health state classification. Adv. Eng. Inform. **32**, 139–151 (2017)
18. W. Mao, Y. Liu, L. Ding, Imbalanced fault diagnosis of rolling bearing based on generative adversarial network: a comparative study. IEEE Access. **7**, 9515–9530 (2019)
19. M. Nuthal Srinivasan, M. Chinnadurai, An efficient video inpainting approach using deep belief network. Comput. Syst. Sci. Eng. **43**(2), 515–529 (2022)
20. L. Qing, H. Rui, D. Xiangqian, Fault diagnosis of rolling bearing based on improved stacking self encoder. Comput. Eng. Des. **40**(7), 2064–2070 (2019)
21. Y. Ren, J. Bai, J. Zhang, Label contrastive coding based graph neural network for graph classification, in *Int. Conf. on Database Systems for Advanced Applications*, China (2021) pp. 123–140
22. F. Scarselli, M. Gori, A.C. Tsoi, The graph neural network model[J]. IEEE Trans. Neural Netw. **20**(1), 61–80 (2008)
23. H. Shao, H. Jiang, Electric locomotive bearing fault diagnosis convolutional based on deep belief network and multisensory information fusion. IEEE Trans. Industr. Electron. **65**(3), 2727–2736 (2017)

24. M. Sharma, R.B. Pachori, U.R. Acharya, A new approach to characterize epileptic seizures using analytic time-frequency flexible wavelet transform and fractal dimension. Pattern Recogn. Lett. **94**(7), 172–179 (2017)
25. H.D. Shao, H.K. Jiang, H.Z. Zhang, Rolling bearing fault feature learning using improved convolutional deep belief network with compressed sensing. Mech. Syst. Signal Process. **100**, 743765 (2018)
26. F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner, S. Mushtaq, M. Islam, M. Sohaib, Deep learning aided data-driven fault diagnosis of rotatory machine. A comprehensive review. Energies **14**(16), 5150 (2021). https://doi.org/10.3390/en14165150
27. A.A. Salamai, A.A. Ageeli, E.M. El-kenawy, Forecasting e-commerce adoption based on bidirectional recurrent neural networks. Comput. Mater. Contin. **70**(3), 5091–5106 (2022)
28. N. Upadhyay, J. Metsebo, P.K. Kankar et al., An improved theoretical model of unbalanced shaft-bearing system for accurate performance prediction of ball bearing due to localized defects. Iran J. Sci. Technol. Trans. Mech. Eng. **42**, 293–309 (2018). https://doi.org/10.1007/s40997-017-0098-9
29. O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, Matching networks for one shot learning. Adv. Neural. Inf. Process. Syst. **34**(3), 3630–3638 (2016)
30. G. Victor, B. Joan, Few-shot learning with graph neural networks, Preprint at http://arxiv.org/abs/1711.04043 (2017)
31. Y. Wang, G. Xu, L. Lang, Detection of weak transient signals based on wavelet packet transform and manifold learning for rolling element bearing fault diagnosis. Mech. Syst. Signal Process. **54**, 259–276 (2015)
32. J. Wang, Z. Fang, N. Lang, A multi-resolution approach for spinal metastasis detection using deep Siamese neural networks. Comput. Biol. Med. **84**, 137–146 (2017)
33. Z. Wei, C. Li, G. Peng, A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. Mech. Syst. Signal Process. **100**, 0439–0453 (2018)
34. Z. Wang, J. Wang, Y. Wang, An intelligent diagnosis scheme based on generative adversarial learning deep neural networks and its application to planetary gearbox fault pattern recognition. Neurocomputing **310**, 213–222 (2018)
35. X. Zhou, l, Jiang Z H, Gear fault diagnosis based on improved HHT and Mahalanobis distance. J. Vib. Shock **36**(22), 218–224 (2017)
36. B. Zhang, W. Li, J. Hao, Adversarial adaptive 1-Dconvolutional neural networks for bearing fault diagnosis under varying working condition. Preprint at http://arxiv.org/abs/1805.00778 (2018)
37. J. Zhang, R.X. Gao, Deep learning-driven data curation and model interpretation for smart manufacturing. China Mech. Eng. (2021). https://doi.org/10.1186/s10033-021-00587-yl
38. J. Zhao, S. Gao, Y. Liu, Anomaly detection and pattern differentiation in monitoring data from power transformers. Energy Eng. **119**(5), 1811–1828 (2022)

Publisher's Note