Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH

Open Access

Bearing fault diagnosis method based on improved Siamese neural network with small sample

Xiaoping Zhao[1,2], Mengyao Ma[1*] and Fan Shao[3]

Abstract

Fault diagnosis of rolling bearings is very important for monitoring the health of rotating machinery. However, in actual industrial production, owing to the constraints of conditions and costs, only a small number of bearing fault samples can be obtained, which leads to an unsatisfactory effect of traditional fault diagnosis models based on data-driven methods. Therefore, this study proposes a small-sample bearing fault diagnosis method based on an improved Siamese neural network (ISNN). This method adds a classification branch to the standard Siamese network and replaces the common Euclidean distance measurement with a network measurement. The model includes three networks: a feature extraction network, a relationship measurement network, and a fault classification network. First, the fault samples were input into the same feature extraction network in pairs, and a long and short-term memory (LSTM) network and convolutional neural network (CNN) were used to map the bearing signal data to the low-dimensional feature space. Then, the extracted sample features were measured for similarity by the relationship measurement network; at the same time, the features were input into the classification network to complete the bearing fault recognition. When the number of training samples was particularly small (training set A, 10 samples), the accuracy of 1D CNN, Prototype net and Siamese net were 49.8%, 60.2% and 58.6% respectively, while the accuracy of the proposed ISNN method was 84.1%. For the 100-sample case of training set D, the accuracy of 1D CNN was improved to 93.4%, which was still higher than that of prototype and Siam network, while the accuracy of ISNN method reached 98.1%.The experimental results show that the method in this study achieved higher fault diagnosis accuracy and better generalization in the case of small samples.

Keywords: Rolling bearing, Siamese neural network, Small sample, Fault diagnosis

Introduction

Rolling bearings are key components in rotating machinery. They are widely used in aerospace, rail transit, industrial production, and other fields. Once they fail, they directly affect the normal operation of the entire equipment and cause economic losses to enterprises or even lead to accidents that threaten people's lives and safety. Therefore, it is highly desirable to accurately identify the fault status of rolling bearings to monitor the health of mechanical equipment and eliminate potential safety hazards in time.

The rolling bearing fault diagnosis methods in the early days mostly used signal decomposition and transformation technology to extract fault features manually, such as empirical mode decomposition [1] and wavelet packet transform [2]. In recent years, machine learning and deep neural network methods based on big data have led to rapid developments in many tasks such as target detection [3], semantic segmentation [4], and image classification [5]. In the field of fault diagnosis, an increasing number of scholars have applied the following methods

*Correspondence: 20211249492@nuist.edu.cn

[1] School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China
Full list of author information is available at the end of the article

to monitor the health condition of rolling bearings [6]: support vector machines (SVM) [7, 8], long short-term memory (LSTM) networks [9], and convolutional neural networks (CNN) [10, 11]. Although the above methods have been able to achieve good fault diagnosis results, they over rely on large-scale and high-quality training data. In actual practice, the working condition environment of a rolling bearing is complex; it is unrealistic to shut equipment down frequently to collect fault signals, and the faulty equipment may not be able to continue running; as a result, only limited fault data can be collected if any at all. In other words, it is difficult or expensive to collect a large amount of labeled fault data, which leads to a small-sample fault diagnosis problem [12]. In this case, deep learning models are unsatisfactory (low accuracy and poor generalization) and result in a serious overfitting phenomenon. Therefore, the study of small-sample fault diagnosis methods cannot only realize the accurate identification of equipment health status under limited training data, but also be of great significance in alleviating the difficulty of fault signal acquisition and reducing the investment of human and material resources.

The theory of small-sample learning [13] has attracted extensive research in recent years. For the problem of small-sample recognition in various fields, researchers have proposed many excellent methods that can be classified as data enhancement, transfer learning, meta learning, and metric learning [14]. Many studies on small-sample fault diagnosis have been published in the field of mechanical fault diagnosis. Wang et al. [15] enhanced the quality of generation samples and the ability of planetary gearbox fault diagnosis by combining a generative adversarial network and stacked denoising autoencoders. Lv et al. [16] proposed a semi-supervised fault diagnosis method based on the augmentation of a gearbox-labeled sample in a deep embedding relation space that improved the generalization ability of the relation network by expanding the labeled samples and realized gearbox fault identification under a small number of labeled samples. Hu et al. [17] used order tracking and resampling methods to process bearing data at different speeds, and an adaptive batch standardized network was applied to classify cross-working condition faults in small samples. Chen et al. [18] extended the least square support vector machine to implement a transfer learning strategy and achieved a better diagnostic performance for the bearing fault diagnosis of insufficient labeled samples. Wu et al. [19], Xu et al. [20], and Zheng et al. [21] proposed a small-sample bearing fault diagnosis method based on transfer learning that transfers the knowledge learned from the source domain to the target domain; thus, a good diagnostic accuracy was achieved with a small amount of target data. Regarding metric learning, Zhang et al. [22] proposed a bearing fault diagnosis method based on a Siamese neural network under the condition of small samples that learned features by exploiting sample pairs of the same or different categories. The experimental results showed that the proposed method was more effective for fault diagnosis with limited data. In addition, Wang et al. [23] applied an improved prototypical network to a classification problem based on metric space. Using standard bearing fault data set, its higher performance under the conditions of limited samples and strong noise was verified. Although the above methods improved the fault diagnosis performance for small samples to a certain extent, there are still many problems. For example, the data enhancement method may generate noisy data, and the generation model is often difficult to train. The transfer learning method requires a large amount of source domain data, and the transfer effect depends on the similarity between the target and source domains. In addition, the choice of transfer strategy is very important. Because the meta learning method is highly complex and related technology is not mature, the application of this method is limited. Metric learning only uses simple distance measurements as training guidelines and has low accuracy when there are few training samples, but it is relatively popular because of its simple calculation and easy operation.

The Siamese network [24] is a small-sample learning method based on similarity measurements. It has achieved significant effects in the fields of visual tracking [25], speech processing [26], and signature verification [27]. However, in many cases, the model performance depends on the quality of the feature and the choice of the metric method. When testing the model, every training sample needs to be paired with a testing sample one by one to calculate the similarity. In view of this, aiming at the problem of low accuracy and over-fitting of fault diagnosis by the deep neural network method under the condition of small samples, a rolling bearing fault diagnosis algorithm based on an improved Siamese neural network (ISNN) is proposed in this study. Compared with previous work on bearing fault diagnosis, the main advantages of the proposed method are as follows:

(1) classification branch is added to the standard Siamese network so that the model not only calculates the sample similarity but also predicts and classifies the samples directly, thus avoiding the pairing calculation of the data one by one during the model test. In addition, the standard Siamese network only uses the similarity label information between samples, but the classification branch also effectively

uses the category label information of each sample and can perform a better constraint role in model training.

(2) During feature extraction, the time domain data and frequency domain data are jointly input into the model, and the temporal and spatial features of fault signals are extracted by an LSTM network and CNN to make full use of the information of limited samples.

(3) The fixed distance measurement method is replaced by neural network measurement so that the model can adaptively adjust the measurement method according to the learned characteristics. At the same time, to reduce the number of parameters and alleviate overfitting of the model, a global average pooling layer is used in the measurement and classification networks.

(4) Compared with other methods, the ISNN has a higher diagnostic accuracy under the condition of small samples and also has better generalization under variable working conditions.

The remainder of this paper is organized as follows. Improved siamese neural network section introduces the ISNN and mainly includes the metric learning theory, Siamese network theory, and the proposed ISNN method. Bearing fault diagnosis process based on ISNN section presents the proposed bearing fault diagnosis framework based on the ISNN method. In Experiment and analysis section, the effectiveness of the proposed method is demonstrated, including a dataset introduction and comparison effect analysis. Finally, Conclusion section summarizes the full text.

Improved siamese neural network

Metric learning

Metric learning, also known as similarity learning, refers to the calculation of the distance between two samples through a given distance function to measure their similarity [28]. The goal is to decrease the distance between samples of the same type in an embedding space and to increase the distance between samples of different types. When performing small-sample classification, the final classification result is determined by calculating the distance between the sample to be tested and the sample with a known label and finding the nearest neighbor category. In practical applications, the Euclidean distance and cosine similarity are usually used as distance functions. Representative methods are prototypical networks [29], matching networks [30], and Siamese networks.

Given two samples, x_i and x_j, $f(\cdot)$ represents the feature mapping of the samples, and their Euclidean distance in the metric space can be described as Eq. (1):

$$d_f(x_i, x_j) = d(f(x_i) - f(x_j)) = \left\| f(x_i) - f(x_j) \right\|_2 \quad (1)$$

The ultimate goal of metric learning is to learn an appropriate mapping function under certain constraints.

Siamese neural network

A Siamese neural network is a type of metric learning that solves the classification problem in the case of few samples by measuring the similarity between two samples. This method uses two weight-sharing subnetworks to receive two input samples simultaneously, and the output result is the similarity between the two samples [31]. By pairing training samples into the model, the number of training times can be effectively increased and the relationship between various samples can be deeply explored. First, the model maps the two input samples (X_1,X_2) to the low-dimensional feature space and then calculates the Euclidean distance between the two feature vectors. The distance is used to measure the similarity between samples. A large distance represents a low similarity, whereas a small distance indicates high similarity. The structure of a Siamese network is shown in Fig. 1.

As shown in Fig. 1, the input of the Siamese neural network is a pair of samples, and the output is the similarity between them. When two samples belong to the same category, the similarity value approaches 1; when the two samples belong to different categories, the similarity value approaches 0. The Siamese neural network uses the contrast loss function to optimize the training target of the model as Eq. (2):

$$\begin{aligned} L\left(\mathbf{w}, \left(x_1, x_2, y\right)\right) &= \frac{1}{2} y \left\| f(x_1) - f(x_2) \right\|_2^2 \\ &+ \frac{1}{2}\left(1 - y\right) \max\left(m - \left\| f(x_1) - f(x_2) \right\|_2, 0\right)^2 \end{aligned} \quad (2)$$

where x represents the input sample, y is the similarity label of the sample, and $y = 1$ indicates that the two samples are similar. If the Euclidean distance in the feature space is large, it indicates that the current model does not perform well and increases the loss. $y = 0$ indicates that the two samples are not similar. If the Euclidean distance between the two samples in the feature space is small, the loss value will also increase.m is the set threshold, N is the number of samples, and $\|\cdot\|_2$ is the two-norm between the features, that is, the Euclidean distance.

Introduction to ISNN network structure

The standard Siamese network uses Euclidean distance as the measurement function. The measurement effect depends on the quality of feature extraction in the early stage, and a cumbersome sample comparison is required during model testing. Because the algorithm in this study was performed under the condition

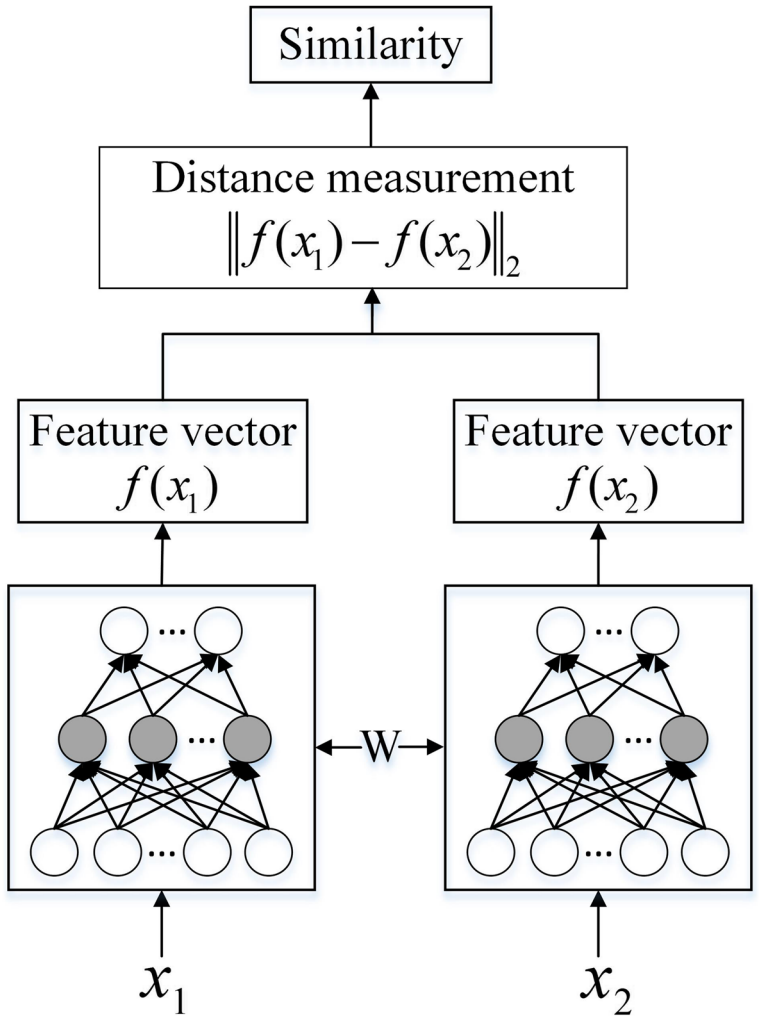

Fig. 1 Structure of Siamese network.The input of the Siamese neural network is a pair of samples, and the output is the similarity between them. When two samples belong to the same category, the similarity value approaches 1; when the two samples belong to different categories, the similarity value approaches 0

of small samples, it was necessary to make full use of the information of each training sample, and the network design was not too deep. To make the model more flexible for fault classification and make full use of the limited sample information, this study added a classification branch to the standard Siamese neural network and redesigned the feature extraction and distance measurement functions according to the characteristics of the bearing data; thus, the proposed ISNN bearing fault diagnosis model included three networks: feature extraction, relationship measurement, and fault classification. The structure is shown in Fig. 2.

Feature extraction network

During data preprocessing, the time domain data and frequency domain data of each training sample were spliced, and the output was used as the input for the feature extraction network. The constructed feature extraction network consisted of two submodules with exactly the same structure and parameters. Each submodule first used two LSTM layers (L1 and L2 in Fig. 2) to extract the time information of the fault sample and then extracted the spatial information through the convolutional layer C1. Finally, a maximum pooling layer P1 was connected to down-sample the convolution result to reduce the feature size. In addition,

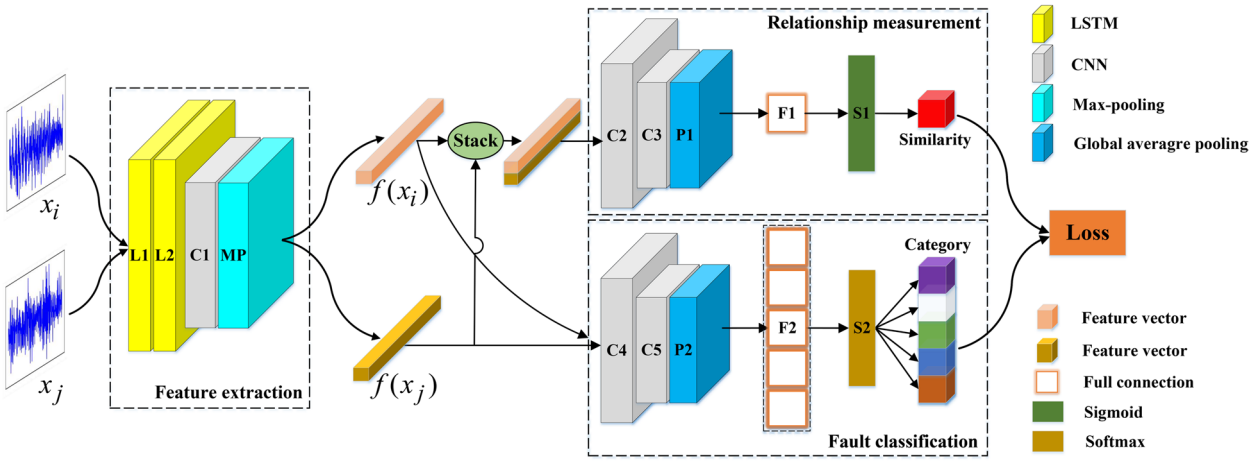

Fig. 2 Structure diagram of ISNN model.The proposed ISNN bearing fault diagnosis model included three networks: feature extraction, relationship measurement, and fault classifications

the training samples were input into the ISNN model in pairs. Assuming that there were n fault samples, every time a pair of samples is input to the model, a total of C_n^2 times effective training can be performed, which greatly increases the number of training times of the model.

Relationship measurement network

The function of the relationship measurement network is to map the two feature vectors $f(X_i)$ and $f(X_j)$ to similarity probability. When two samples are similar, the output probability is 1, and when two samples are not similar, the output probability is 0. Commonly fixed measurement methods overly rely on the spatial quality of the feature embedding learned by a feature extraction network. This study used neural networks to measure the similarity relationship between features, trained it jointly with the feature extraction network, and adaptively adjusted the measurement method according to the input features. As can be seen from the ISNN model structure in Fig. 2, the relationship measurement network was composed of convolutional layers C2 and C3, a global average pooling layer P2, and a fully connection layer F1. Global average pooling was used to replace the entire feature map with its average value. A large number of model parameters were reduced. First, the feature vectors $f(X_i)$ and $f(X_j)$ were put into the convolutional layer, then they were mapped to a similarity value through the P2 and F1 layers, and finally the activation function layer S1 was used to transform the similarity value into [0, 1] . The similarity value was calculated as Eq. (3).

$$R_{i,j} = Sigmoid\left(g\left(f(X_i), f\left(X_j\right)\right)\right) \tag{3}$$

where, $R_{i,j}$ represents the similarity value between the $i-th$ sample and the $j-th$ sample, $g(\cdot)$ denotes the relationship function that maps the feature vector to the similarity value, $f(\cdot)$ represents the output of the feature extraction network, and *Sigmoid* is the activation function.

To accurately measure the similarity between samples of easily confused categories, a weighted similarity loss function was defined. A penalty coefficient was added according to the degree of difficulty in distinguishing between different fault categories to increase the misjudgment loss among easily confused fault categories. The loss function is given as Eq. (4).

$$L_S = \sum_{i,j=1,i\neq j}^{N} \left[Y_{i,j}\left(1 - R_{i,j}\right)^2 + \alpha_{i,j}\left(1 - Y_{i,j}\right)\left(0 - R_{i,j}\right)^2 \right] \tag{4}$$

where L_S is the similarity loss, $Y_{i,j}$ represents the similarity label between two samples, and $\alpha_{i,j}$ denotes the penalty coefficient when samples i and j belong to different fault types. When the two samples are not similar, if the $R_{i,j}$ value output by the network does not approach zero, a larger $\alpha_{i,j}$ value is assigned to increase the loss.

Fault classification network

Because the relationship measurement network only uses the similarity label of the sample and can only judge the similarity of a pair of samples, it cannot directly classify the testing sample; thus, we introduced a classification network in the ISNN model. The designed fault classification network used the category label information of each sample for supervised learning to directly predict the fault category to which it belonged to increase the flexibility of the model. As shown in Fig. 2, the structure of the fault classification network was similar to that of the relationship measurement network consisting of

convolutional layers C4 and C5, a global average pooling layer P3, and a fully connected layer F2. At the end of the network, an activation function layer S2 was used to output the predicted probability of each fault category. During network training, two training samples were separately input to the fault classification network for prediction, but only one sample needed to be input in the testing phase. In this section, we used the mean square error as the loss function as Eq. (5).

$$L_C = \frac{1}{2N}\sum_{i,j=0}^{N}\left[\left(Y_i - y_i\right)^2 + \left(Y_j - y_j\right)^2\right] \tag{5}$$

where L_C is the classification loss, $Y_{(\cdot)}$ represents the real category label of the sample, and $y_{(\cdot)}$ represents the predicted label of the network. The calculation method is Eq. (6).

$$y = Softmax\left(h\left(f(X)\right)\right) \tag{6}$$

where $h(\cdot)$ denotes the output of the classification network, and *Softmax* is the activation function.

In the ISNN fault diagnosis model, the feature extraction network performs preliminary feature extraction on the input samples (X_i, X_j). The relationship measurement network uses similarity information to constrain the network training such that the feature distance of similar samples becomes closer and the feature distance of samples belonging to different categories becomes farther. The classification network completed the final fault classification task. The three parts were mutually constrained and made full use of the time domain, frequency domain, label, and sample similarity information of the fault data under the condition of small samples. In this study, the metric learning concept was applied to the classification problem, and the entire model adopted a relatively shallow network structure that effectively controlled the size of the parameter. When training the model, the similarity loss L_S and L_C the classification loss were optimized simultaneously, and the two losses were combined to obtain the final loss function of the model as Eq. (7).

$$Loss = L_S + L_C \tag{7}$$

Bearing fault diagnosis process based on ISNN

Based on the proposed ISNN model, this study designed a rolling bearing fault diagnosis process that included three steps: data preprocessing, ISNN training, and fault diagnosis as shown in Fig. 3.

(1) Data Preprocessing. In this study, an acceleration sensor was used to collect the vibration signals of rolling bearings from the fault diagnosis experiment platform. The first half of the collected signal was used as the training set (red box in Fig. 4), and the second half was used as the testing set (green box in Fig. 4). The signal was then divided into many segments according to every 2000 points, and the corresponding frequency domain data were obtained by conducting fast Fourier transform (FFT) on each segment of the signal and concatenating the data before and after the transformation to obtain each sample. Finally, to better perform subsequent model training, the input samples were processed into standardized data of the same magnitude, and each sample was standardized as Eq. (8).

$$X = \{x_i\} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{8}$$

(2) ISNN Training. First, the parameters of the ISNN model were initialized. Two fault samples were randomly selected from the training set to form a sample pair as the model input and gradually train the feature extraction, relationship measurement, and fault classification networks. The loss function was minimized through repeated iterations, and a backpropagation algorithm was used to continuously optimize the model parameters.When the maximum training times were reached, the model parameters were saved.

(3) Fault Diagnosis. When performing bearing fault diagnosis, the testing samples were input into the trained feature extraction network to obtain low-dimensional feature vectors, and then the fault diagnosis results were output by the classification network.

Experiment and analysis

Experimental data

To verify the specific superiority of the proposed ISNN algorithm, five types of bearing vibration signals on different health states were collected from the drivetrain diagnostic simulator as experimental data. The five different fault types were normal state, ball fault, inner fault, outer fault, and combined fault (ball, inner, and outer faults). The structure of the experimental platform is shown in Fig. 5 and was mainly composed of a drive

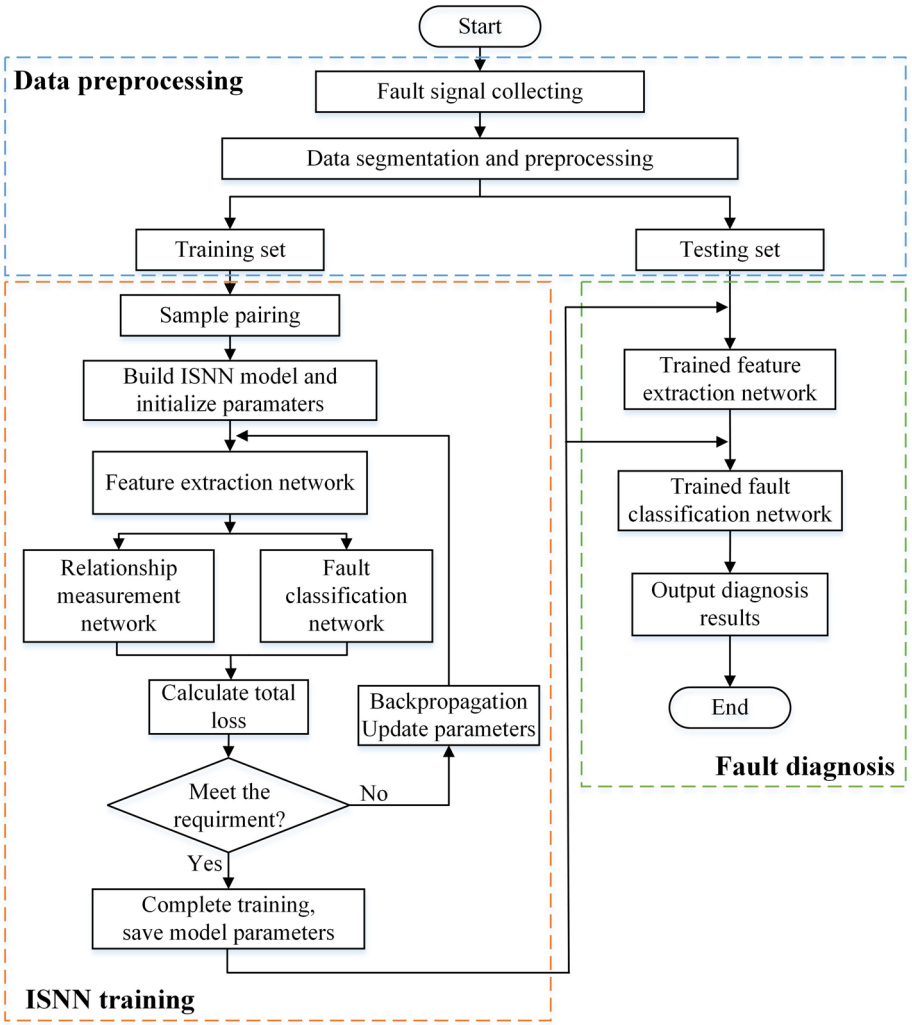

Fig. 3 Bearing fault diagnosis process.Based on the proposed ISNN model, this study designed a rolling bearing fault diagnosis process that included three steps: data preprocessing, ISNN training, and fault diagnosis

motor, gear box, magnetic powder brake, and adjustable load. Figure 6 shows the four types of bearing faults. The letters marked in the red box indicate the fault type of each bearing. Figure 7 shows the time domain vibration waveforms of each bearing fault.

In this study, we aimed to prove through simulation the difficulty of collecting abundant fault data under various working conditions in engineering practice by collecting bearing fault signals with only one working condition as training data. The motor speed was set to 1700 r/min, the load voltage was 4 V, and a unidirectional acceleration sensor was used to collect the vibration signal. A sampling frequency of 20 kHz and sampling duration of 20 s was chosen. The vibration signal file of each fault bearing type contained a total of 409600 points. We set the first 10 s of data as the training set and the last 10 s of data as the testing set. The collected bearing signals were divided into segments of 2000 points, and 100 training samples and 100 testing samples were obtained for each fault type. To study the influence of different numbers of fault samples on the ISNN model, 10, 20, and 50 samples were randomly selected from 100 training samples of each type to construct four different training sets. In addition, the fault data of two other working conditions were collected to test the generalization performance of the ISNN. One working condition was similar to the working condition of the training data (1700-r/min motor speed, 8-V load voltage), and the other working condition was quite different from the training data (3400-r/min motor speed, 8-V load voltage. The sampling duration of the testing data was 10 s, and the sample segmentation method was the same as before. The datasets used in this study are shown in Table 1.

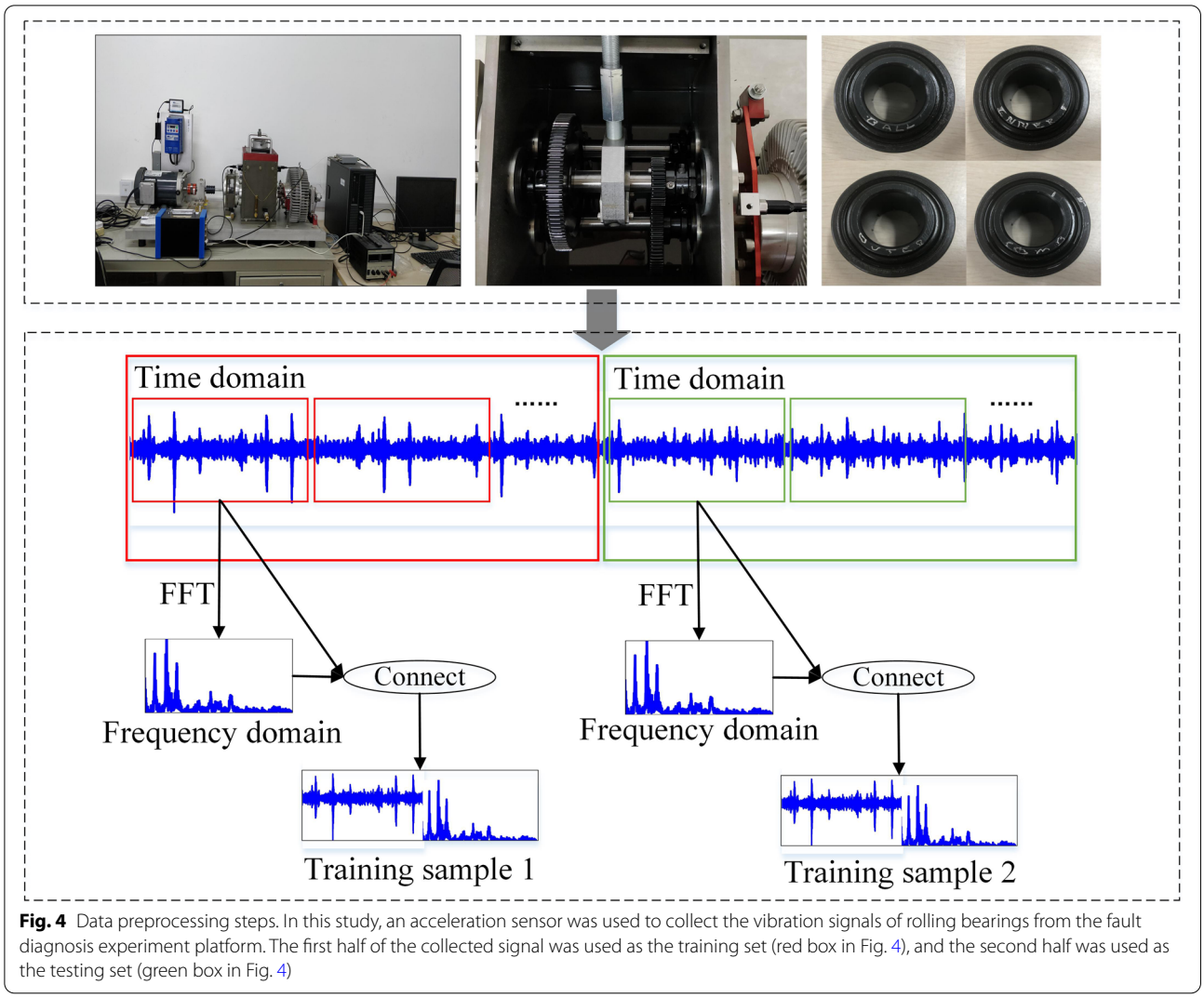

Fig. 4 Data preprocessing steps. In this study, an acceleration sensor was used to collect the vibration signals of rolling bearings from the fault diagnosis experiment platform. The first half of the collected signal was used as the training set (red box in Fig. 4), and the second half was used as the testing set (green box in Fig. 4)

Experimental parameter settings

The experimental software platforms used were i7-4790 CPU, NVIDIA GTX1050Ti, Python3.7, and Pytorch1.3. In the experiment, the Adam algorithm was used to optimize the model parameters, the batch size was set to 50, the learning rate was set to 0.002, and the maximum number of iterations was 400. The detailed parameter settings of the ISNN model are presented in Table 2. Before being input to the model, the fault signals were transformed into a 40×100 size, where 40 represents the single time-sequence input size of the LSTM network, and 100 denotes the total number of input sequences. In Table 2, the parameters of the 1D CNN represent the input channel, output channel, kernel size, step size, and padding size; the pooling layer parameters represent the pooling window size and step size.

Performance analysis of ISNN

Influence of sample size on model performance

To demonstrate the influence of different numbers of training samples on the fault diagnosis performance of the ISNN, the number of training samples for each fault type was set to 10, 20, 50, and 100 in Training sets A, B, C, and D, respectively, and Testing set 1 was used to verify the effect of the model. The accuracy curve of the ISNN is shown in Fig. 8. The red solid line is the accuracy of the training set, the blue solid line is the accuracy of the testing set, and the green dashed line is the reference value. Figure 9 shows the final diagnostic accuracy value of the ISNN using Testing set 1 with the different training sets.

A shown in Fig. 8, the training accuracy of the model with different training set sizes quickly reached 100% (red solid line), particularly for Training sets A and B. This

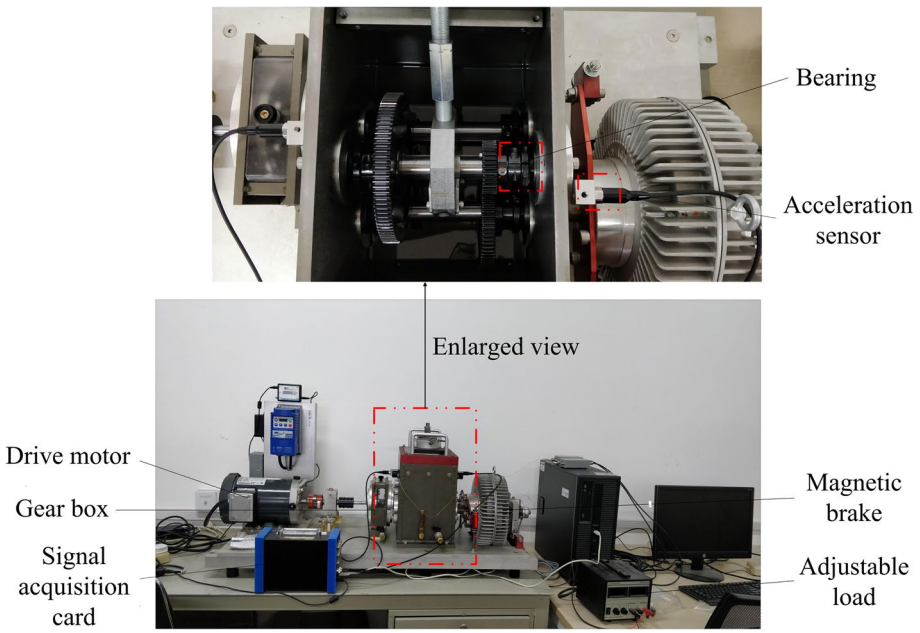

Fig. 5 Drivetrain diagnostic simulator. The structure of the experimental platform was mainly composed of a drive motor, gear box, magnetic powder brake, and adjustable load

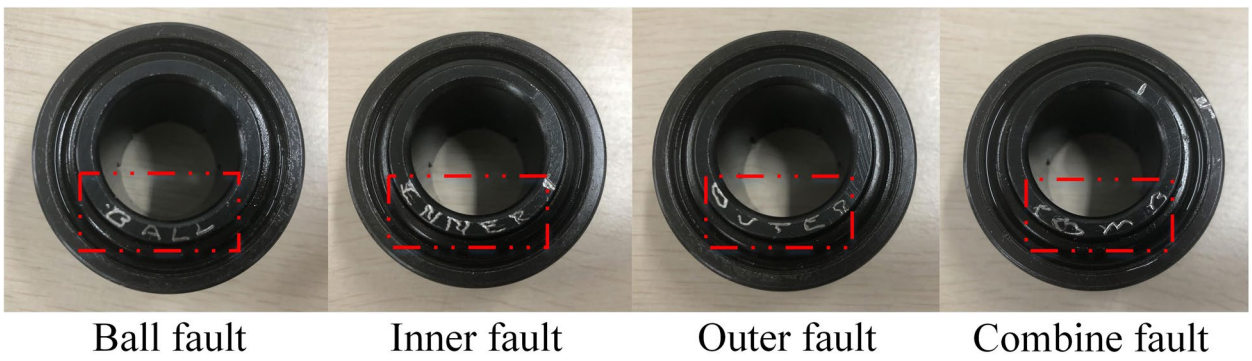

Fig. 6 Bearing fault states. The letters marked in the red box indicate the fault type of each bearing

shows that in the case of small samples, the ISNN learned the information from the training data very easily. However, in the testing set, the diagnostic effect of the model was greatly affected by the number of samples. As shown in Fig. 8(a), with 10 training samples for each fault type, the testing accuracy of the ISNN was approximately 84%, and the model exhibited an overfitting phenomenon. As shown in Fig. 8(b), when 20 training samples were used for each fault type, the testing accuracy exceeded 92%, and the overfitting problem was significantly relieved. As shown in Fig. 8(c), increasing the number of training samples to 50 for each fault type obtained a testing accuracy of approximately 95%, and in Fig. 8(d), the testing accuracy for 100 samples was close to 99% and effectively realized the diagnosis of bearing faults. Moreover, comparing Fig. 8 (a)-(d), the model training process fluctuated significantly when the number of training samples was small. As the number of samples increased, the training process became more stable, and the testing accuracy curve became smoother.

In Fig. 9, it is apparent from the overall results that the diagnostic performance of the ISNN model increased as the number of training samples increased. In particular, when the number of training samples increased from 10 to 20, the testing accuracy increased from 84.1% to 92.3%. The above results show that the ISNN model's dependence on large-scale training data was significantly reduced. By alleviating the overfitting problem effectively

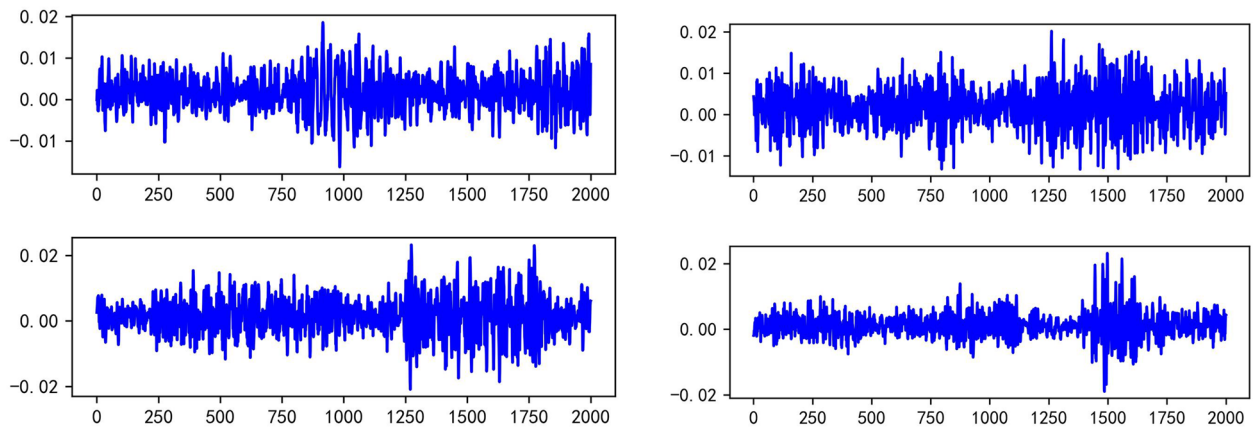

Fig. 7 Vibration waveforms of faulty bearings. This figure shows the time domain vibration waveforms of each bearing fault. (Upper left) (a) Ball fault (Upper right) (b) Combined fault (Lower left) (c) Inner fault (Lower right) (d) Outer fault

Table 1 Experimental datasets

Dataset	Number of samples of each fault type					Working condition (speed, voltage)
	Normal	Ball fault	Combined fault	Inner fault	Outer fault	
Training set A	10	10	10	10	10	1700 r/min, 4 V
Training set B	20	20	20	20	20	
Training set C	50	50	50	50	50	
Training set D	100	100	100	100	100	
Testing set 1	100	100	100	100	100	1700 r/min, 4 V
Testing set 2	100	100	100	100	100	1700 r/min, 8 V
Testing set 3	100	100	100	100	100	3400 r/min, 8 V

Table 2 Parameter setup of ISNN

Network name	Layer type	Main parameters	Output size
Feature extraction network	Input	\	40 × 100
	LSTM1	(40,10)	10 × 100
	LSTM2	(40,10)	10 × 100
	1D CNN	(1,16,9,3,1)	16 × 332
	Max pooling	(5,3)	16 × 110
Relationship measurement network	Input	\	32 × 110
	1D CNN	(32,32,5,2,1)	32 × 54
	1D CNN	(32,16,5,2,1)	16 × 26
	Global average pooling	\	16
	Full connection	(16,1)	1
Fault classification network	Input	\	16 × 110
	1D CNN	(16,32,5,2,1)	32 × 54
	1D CNN	(32,16,5,2,1)	16 × 26
	Global average pooling	\	16
	Fully connected	(16,5)	5

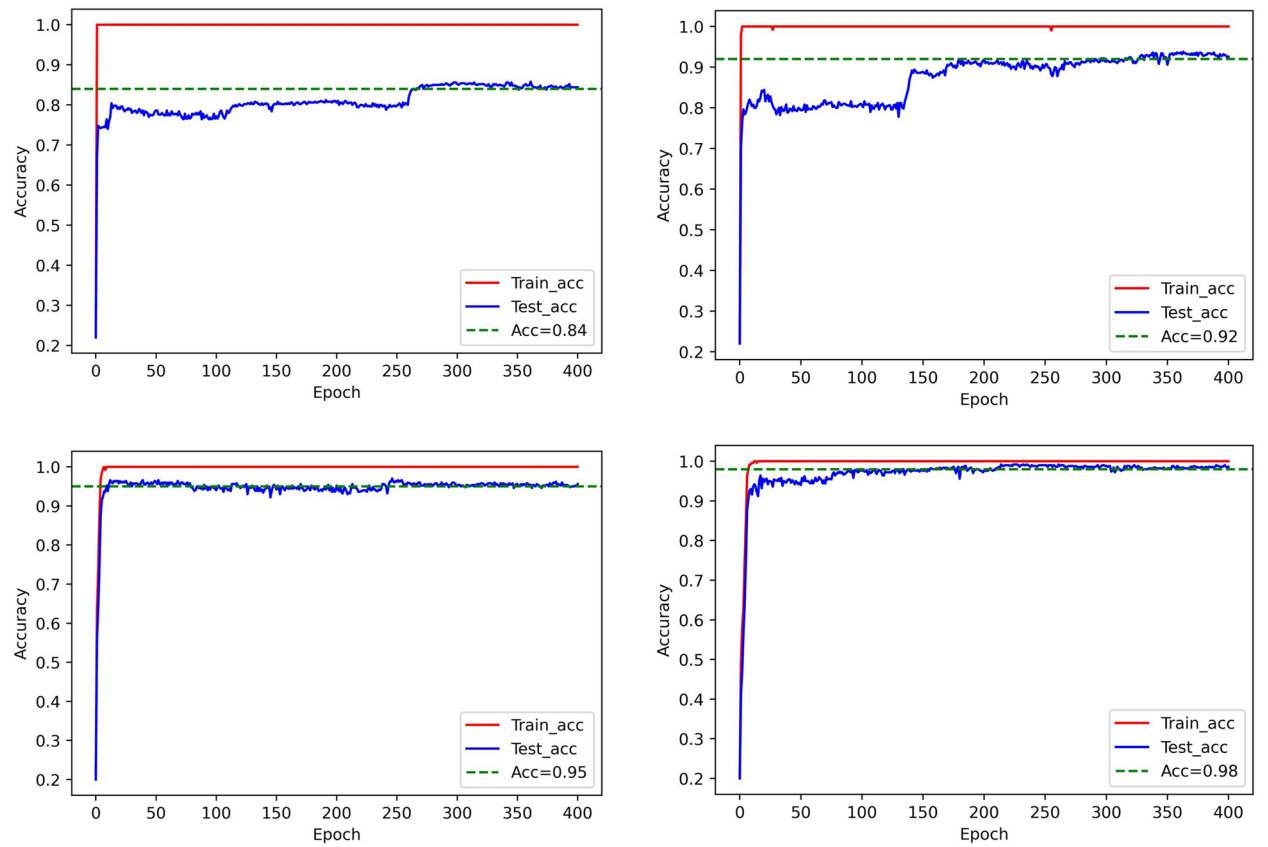

Fig. 8 Training and testing accuracy curves under different number of training samples. (Upper left) (a) Training Set A (Upper right) (b) Training Set B (Lower left) (c) Training Set C (Lower right) (d) Training Set D

in the case of small samples, the ISNN still achieved improved results when there were only dozens of fault samples.

Influence of different feature extraction methods on model performance

To analyze the impact of the feature extraction network of the ISNN model on fault diagnosis performance, four different feature extraction methods were compared. They used only time domain data, frequency domain data, the fully connected layer, and a 1D CNN. The diagnostic accuracy of various feature extraction methods was tested using different training sets, and the results are shown in Table 3 and Fig. 10.

From Table 3 and Fig. 10, we can see that when the time domain data was directly used for model training, the diagnostic result was the worst. For the 10-sample fault types, the accuracy was only 43.4%. When the number of training samples reached 100, the accuracy was only 63.2%. In contrast, when frequency domain data was used for model training, the diagnostic effect was greatly improved. In the case of 10 training samples, the accuracy was over 80%; in the case of 100 training samples, the accuracy reached 95.2%, which was 32% higher than that of the time domain data, indicating that the frequency domain information of the signal provided more effective features for the model. In addition, when using different network structures for feature extraction, if the number of training samples was particularly small, the feature extraction effect of the convolution structure was not as good as that of the fully connected structure. For example, when there were only 10 and 20 training samples, the accuracy of the convolutional structure was 4.8% and 2.2%, respectively, lower than that of the fully connected structure. With an increase in the number of training samples, the feature extraction ability of the convolutional structure improved. In the case of 100 samples, the fault diagnostic accuracy exceeded 97%, indicating that the convolutional structure was dependent on the number of samples. In contrast, the feature extraction network designed in this study with the time domain and frequency domain information input into the model simultaneously and the use of an LSTM network and CNN to

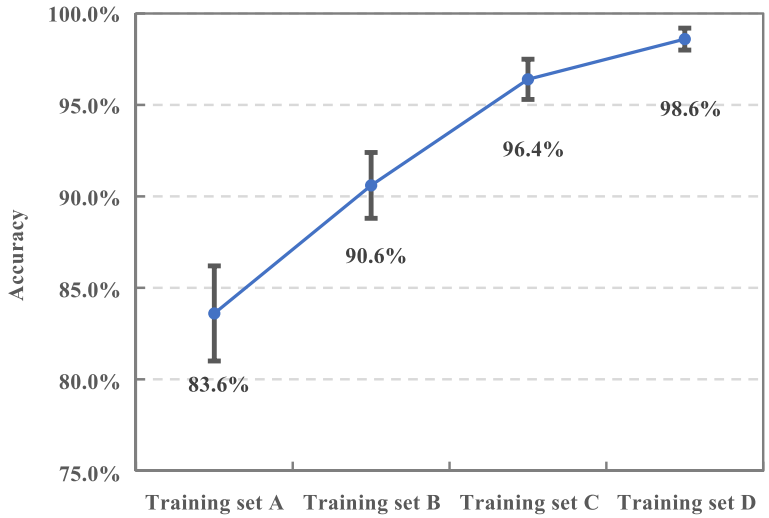

Fig. 9 Testing accuracy of ISNN using different training sets

Table 3 Accuracy of different feature extraction methods

Feature extraction method	Training A	Training B	Training C	Training D
ISNN	84.10%	92.30%	95.00%	98.10%
Time domain	43.40%	48.40%	55.60%	63.20%
Frequency domain	80.20%	87..8%	92.80%	95.20%
Fully connected	77.40%	84.60%	88.00%	93.80%
1D CNN	72.60%	82.40%	91.60%	97.20%

jointly extract sample features achieved the highest fault diagnostic accuracy.

Influence of different relationship measurement methods on model performance

To evaluate the influence of the relationship measurement network of the ISNN model on fault diagnosis performance, under the framework of the algorithm in this study, the measurement method of the proposed ISNN was replaced with the Euclidean distance and cosine distance methods to form two comparison models to verify the effect of the ISNN using different training sets. The results are presented in Table 4 and Fig. 11.

As shown in Table 4 and Fig. 11, under the algorithm framework of this study, the three relationship measurement methods all achieved good diagnostic accuracy, but there were differences under various training sets. In the case of Training set A (10 samples), the cosine distance measurement method had the worst result with an accuracy of only 74.8%, while the accuracy of the proposed ISNN measurement method was 84.1%, which was 2.7% higher than that of the Euclidean distance measurement method. However, as the number of training samples increased, the result of the cosine distance measurement method improved and was eventually better than that of the Euclidean distance measurement method. For Training set D, the diagnostic accuracy of the cosine method reached 97.2% and was higher than that of the Euclidean method with an accuracy of 96.4%. At this time, the accuracy of the ISNN measurement method was 98.1%, which was slightly higher than that of the two fixed distance measurement methods. Overall, when the number of training samples increased from 10 to 100, the accuracy and stability of the ISNN measurement method were better than those of the fixed distance measurement methods.

Performance comparison with other methods

To further investigate the excellent performance of the ISNN model for fault diagnosis with small samples, a 1D CNN, prototype network, and standard Siamese network were selected as comparison methods. The 1D CNN consisted of four convolution modules and three fully connected layers, The prototype network was also

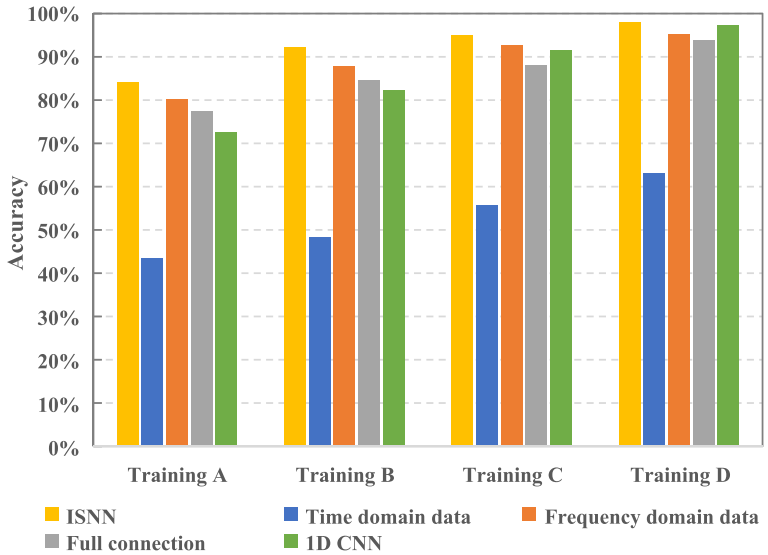

Fig. 10 Accuracy histogram of different feature extraction methods

Table 4 Accuracy of different relationship measurement methods

Measurement method	Training A	Training B	Training C	Training D
Euclidean distance	81.40%	89.40%	94.60%	96.40%
Cosine distance	74.80%	85.80%	95.00%	97.20%
ISNN	84.10%	92.30%	95.00%	98.10%

a small-sample learning method that divided the training set into a support set and a query set to perform the nearest neighbor classification by calculating the distance between a testing sample and a prototype. The standard Siamese network was the small-sample learning method based on the similarity measurement introduced in Siamese neural network section. In the experiment, each model was trained using four training sets. To verify the generalization performance of the ISNN model, three testing sets with different working conditions were used for testing.

Accuracy comparison under same working conditions

Table 5 lists the fault diagnosis accuracy of various methods on Testing set 1 under different numbers of training samples with the same working conditions as the training set. When the number of training samples was particularly small (Training set A), the accuracy of the 1D CNN was the lowest and less than 50%, and the accuracies of the two small-sample prototype and Siamese networks were about 60%. The proposed ISNN method accuracy reached 83.6%. For the 50-sample case of Training set C, the accuracy of the 1D CNN was significantly improved, reaching 88.6%, and was higher than the 82.5% of the prototype network and the 75.8% of the Siamese network. At this time, the accuracy of the ISNN method was still nearly 8% higher than that of the 1D CNN. For the 100-sample case of Training set D, the accuracy of the 1D CNN increased to 93.4%, which was still higher than that of the prototype and Siamese networks, and the gap between the ISNN method became smaller, only 5.2% lower. On the whole, the 1D CNN was significantly affected by the number of training samples. When the sample was severely insufficient, the model became overfitted, and the effect of directly applying it to small-sample fault diagnosis was not ideal. The diagnostic accuracy of the Siamese network before the improvement was very poor and inferior to that of the prototype network. In the case of 100 training samples, the accuracy was only 86.7%. However, the effect of the improved ISNN method increased significantly, obviously better than that of the two comparison methods.

Using Training set B (20 samples) as an example to observe the recognition results of each type of bearing fault by the four methods, the confusion matrix of the diagnosis results of each model using Testing set 1 is shown in Fig. 12. The abscissa represents the predicted fault class, the ordinate represents the real fault class, and the main diagonal represents the number of predicted correct testing samples. There were 100 testing samples for each category.

Figure 12 shows that for 20 training samples for each fault type, the three comparison methods had a large number of misclassifications and the fault diagnostic

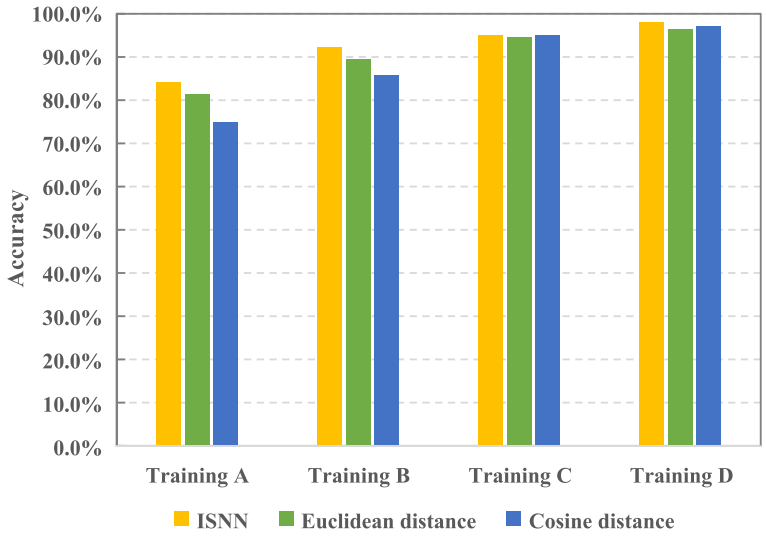

Fig. 11 Accuracy histogram of different relationship measurement methods

effect was very poor, while the proposed ISNN method was obviously better than that of the other methods. As shown in Fig. 12(a), when the 1D CNN was used for diagnosis, only the inner fault and normal samples were classified well, while only 37 of the 100 combined fault samples were correctly predicted. As shown in Fig. 12(b), the classification result of the prototype network was obviously better than that of the 1D CNN. Except for a large number of misclassified outer fault samples (only 43 correctly predicted), the remaining four types of faults all exhibited a certain degree of improvement. As shown in Fig. 12(c), the Siamese network had the worst effect on bearing fault diagnosis. Only the normal samples were identified well (94 samples), while about half of the combined fault and outer fault samples were misclassified. In contrast, Fig. 12(d) shows that the ISNN method significantly improved the classification results of the various fault types. The combined fault case had the worst classification (19 samples incorrectly predicted), and more than 90 samples of the other fault types were correctly identified. From the above analysis, it can be concluded that in the case of a small number of training samples, the several comparison methods ineffectively diagnosed bearing fault, while the proposed ISNN method performed better on small-sample fault diagnosis.

Accuracy comparison under different working conditions

In actual production, the working conditions of equipment often changes, resulting in different distributions of the testing and training data. To verify the generalization performance of the ISNN method compared to other methods on different testing data, Testing set 2 (similar to the training set working condition) and Testing set 3 (largely different from the training set working condition) in Table 1 were used to conduct the experiments, and the results are shown in Table 6.

In Table 6, when the testing data were similar to the training set data, the diagnostic accuracies of the three comparison methods in the 10-sample Training set A and 20-sample Training set B were very low. When the number of training samples was increased to 100 (Training set D), the 1D CNN achieved an accuracy of 90.2%, while the accuracies of the prototype and Siamese networks were only 74.4% and 78.5%, respectively, while the accuracy of the ISNN method was still approximately

Table 5 Accuracy of various methods using Testing set 1

Method	Training set A	Training set B	Training set C	Training set D
1D CNN	49.80%	68.20%	88.60%	93.40%
Prototype net	60.20%	74.60%	82.50%	90.30%
Siamese net	58.60%	66.80%	75.80%	86.70%
ISNN	84.10%	92.30%	95.00%	98.10%

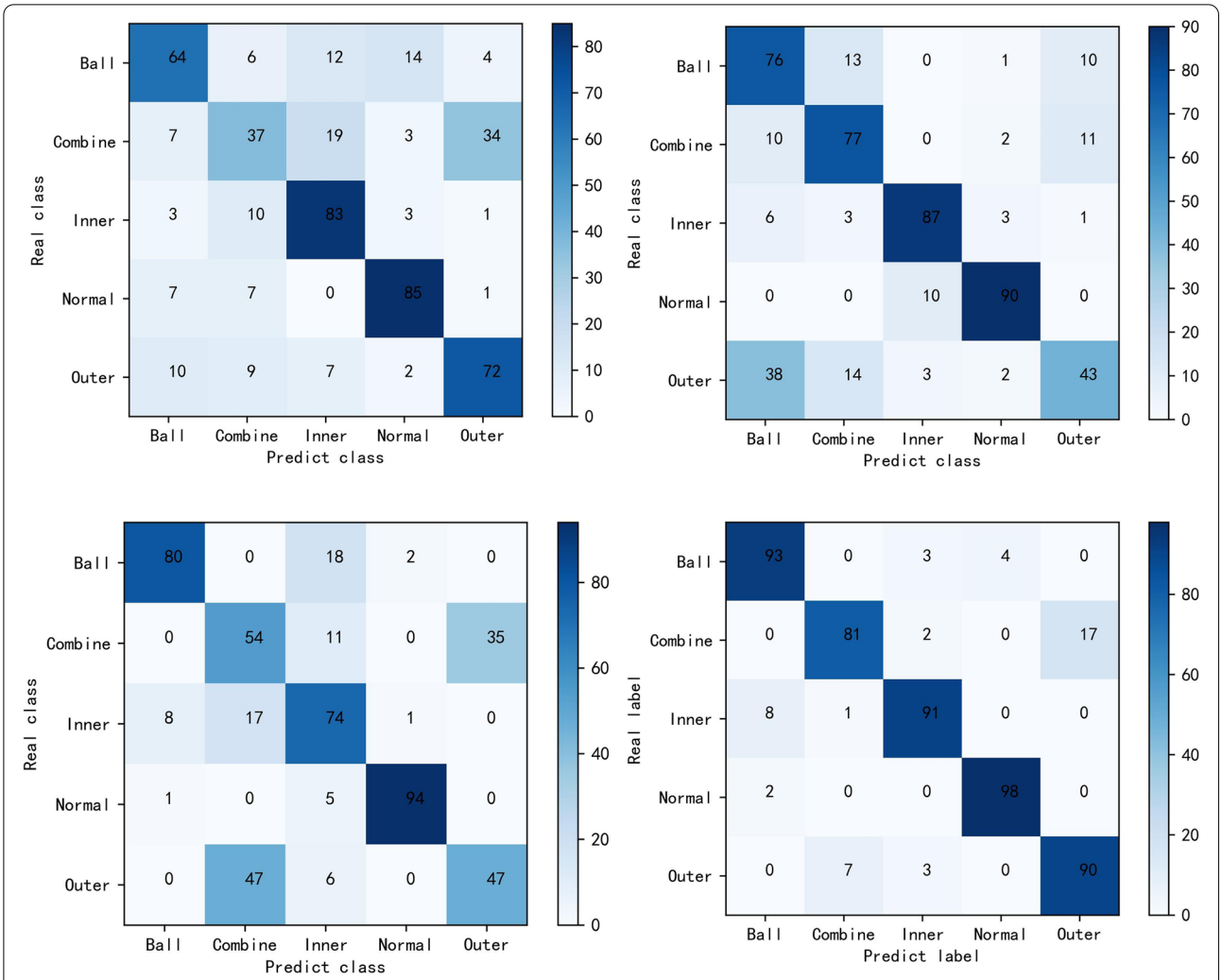

Fig. 12 Confusion matrix of diagnostic results of various methods under Training set B. (Upper left) (a) 1D CNN (Upper right) (b) Prototype network (Lower left) (c) Siamese network (Lower right) (d) ISNN

Table 6 Accuracy of various fault diagnosis methods under different testing sets

Dataset	Testing set 2				Testing set 3			
	1D CNN	Prototype net	Siamese net	ISNN	1D CNN	Prototype net	Siamese net	ISNN
Training set A	49.20%	42.30%	46.40%	83.00%	47.30%	40.40%	45.10%	82.40%
Training set B	67.00%	54.40%	58.30%	89.70%	65.20%	51.60%	55.20%	88.70%
Training set C	85.60%	65.80%	69.60%	94.80%	83.50%	60.90%	66.00%	94.40%
Training set D	90.20%	74.40%	78.50%	97.80%	86.00%	68.60%	76.60%	95.20%

98%. When the working conditions of Testing set 3 were quite different from the training data, the fault diagnosis effect of the comparison methods became worse, particularly when there were less than 50 training samples. The diagnostic accuracy of the prototype and Siamese networks was only approximately 60%, and the 1D CNN had an accuracy of approximately 80%, while the ISNN method achieved an accuracy of more than 90%. From the analysis above, it can be concluded that the prototype and Siamese networks had poor diagnostic accuracy and weak generalization ability when facing variable working condition data. The accuracy of the 1D CNN

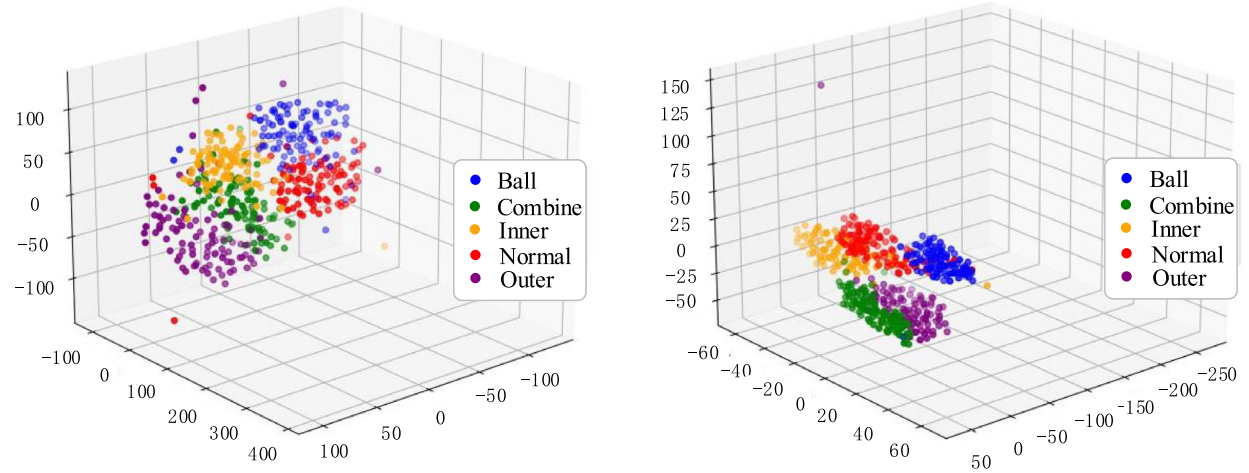

Fig. 13 Visualization of diagnostic results of 1D CNN and ISNN under Training set C. (Left) (a) 1D CNN (Right) (b) ISNN

method also decreased significantly, while the ISNN method was only reduced by approximately 1%, and the highest accuracy of 95.2% was achieved when the working conditions changed significantly. Compared with the other three methods, the generalization performance was significantly improved.

To more intuitively show the fault diagnostic effect of the ISNN method on variable working condition data with small samples, Training set C (50 samples in each category) and Testing set 3 (largely different from the training set working condition), we selected the 1D CNN method as a comparison, and the fault classification results of the two methods were visualized by t-SNE. The results are shown in Fig. 13.

As shown in Fig. 13(a), although the 1D CNN method roughly classified the various fault samples, there were still different degrees of overlap. For example, some samples of the inner fault (yellow scatter) were mixed with the combined fault samples (green scatter), and the normal samples (red scatter) were also partially mixed with the ball fault samples (blue scatter). In addition, a large number of outer fault samples (purple scatter) and combined fault samples (green scatter) overlapped, and the overall classification result of the model was poor. As shown in Fig. 13(b), after the ISNN method was used for fault diagnosis, the grouping of samples of the same fault category improved, and there were larger intervals between the samples of different categories with only a small number of outer fault samples (purple scatter) misclassified as combined faults (green scatter). This indicates that the ISNN method identified various bearing faults well under variable working conditions, verifying that the model has a better generalization effect under small-sample conditions.

Conclusions

In this study, we reported a rolling bearing fault diagnosis algorithm based on the condition of small samples that integrated the metric learning idea and an ordinary classification network into a framework. We achieved an accurate diagnosis of bearing faults by jointly training the feature extraction, relationship measurement, and fault classification networks. The following conclusions can be drawn from the experimental analysis.

(1) The ISNN model effectively diagnoses bearing faults in a variety of small-sample conditions, achieving an accuracy of 84.1% using only 10 training samples and 98.1% using 100 training samples.
(2) The LSTM+CNN feature extraction network designed in this study used neural networks as a relationship measurement method and enabled the model to learn more discriminative sample features, thereby obtaining greater diagnostic accuracy.
(3) Through an experimental comparison with the 1D CNN, prototype network, and original Siamese network methods, the ISNN method proposed in this study had the highest diagnostic accuracy of 95.2% for 100 training samples and had better generalization when the working conditions of testing data changed.

The ISNN proposed in this paper achieves good classification effect in the known and single small-sample fault classification task, but it can not distinguish the unknown type of fault and complex fault. Next, we will try to introduce unknown, complex fault and other fault types on the basis of this paper to more comprehensively verify the superiority of this method.

Authors' contributions
Xiaoping Zhao, Mengyao Ma, and Fan Shao conceived and designed the study.Mengyao Ma and Fan Shao performed the simulations and wrote the paper. Xiaoping Zhao, Mengyao Ma, and Fan Shao conduct the experiment and confirm the conclusion. All authors reviewed and edited the manuscript. All authors read and approved the final manuscript.

Authors' information
Xiao-Ping Zhao, born in 1977, is currently an associate professor at Nanjing University of Information Science and Technology, China. She received her doctor degree from Nanjing University of Aeronautics and Astronautics, China, in 2009. Her research interests include signal processing and fault diagnosis. E-mail: zxp@nuist.edu.cn
Meng-Yao Ma, born in 1999, is currently a master candidate at School of computer and software, Nanjing University of Information Science and Technology, China. Her research interests include deep learning, fault diagnosis. E-mail: 20211249492@nuist.edu.cn
Fan Shao, born in 1997, obtained a master's degree from School of Automation, Nanjing University of Information Science and Technology, China. His research interests include deep learning, fault diagnosis. E-mail: shaofan1206@qq.com

Funding
This research is supported financially by National Natural Science Foundation of China (Grant No.51505234,51575283).

Availability of data and materials
The raw data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study.

Declarations

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

Author details
[1]School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China. [2]Engineering Research Center of Digital Forensics, Nanjing University of Information Science and Technology, Nanjing, China. [3]School of Automation, Nanjing University of Information Science and Technology, Nanjing, China.

Received: 20 July 2022 Accepted: 29 September 2022
Published online: 19 November 2022

References

1. Li Y, Xu M, Huang W, Zuo MJ, Liu L (2017) An improved emd method for fault diagnosis of rolling bearing. In: 2016 Prognostics and System Health Management Conference (PHM-Chengdu), Chengdu, China, pp. 1–5. New York: IEEE
2. Yong-Gang XU, Meng ZP, Ming LU (2014) Fault diagnosis method of rolling bearing based on dual-tree complex wavelet packet transform and svm. Journal of Aerospace Power 29(1):67–73
3. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence 39(6):1137–1149
4. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 39(4):640–651
5. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 770–778. New York: IEEE
6. Chen XH, Cheng G, Shan XL, Hu X, Guo Q, Liu HG (2015) Research of weak fault feature information extraction of planetary gear based on ensemble empirical mode decomposition and adaptive stochastic resonance. Measurement 73:55–67
7. Sugumaran V, Ramachandran KI (2011) Effect of number of features on classification of roller bearing faults using SVM and PSVM. Expert Syst Appl 38(4):4088–4096
8. Medina R, Macancela JC, Lucero P, Cabrera D, Sánchez RV, Cerrada M (2020) Gear and bearing fault classification under different load and speed by using Poincaré plot features and SVM. J Intell Manuf 33(4):1031–1055
9. Wang F, Liu X, Deng G, Yu X, Li H, Han Q (2019) Remaining life prediction method for rolling bearing based on the long short-term memory network. Neural Process Lett 50(3):2437–2454
10. Ozcan IH, Devecioglu OC, Ince T, Eren L, Askar M (2021) Enhanced bearing fault detection using multichannel, multilevel 1D CNN classifier. Electr Eng 5(2):1–13
11. Wang L-H, Zhao X-P, Wu J-X, Xie Y-Y, Zhang Y-H (2017) Motor fault diagnosis based on short-time fourier transform and convolutional neural network. Chinese Journal of Mechanical Engineering 30(06):1357–1368
12. Lei YG, Yang B, Du Z, Lv N(2019) Deep transfer diagnosis method for machinery in big data era. J Mech Eng 55(7):1–8
13. Li FF, Member, IEEE, Fergus R, Member S (2006) One-shot learning of object categories. IEEE Trans Pattern Anal Mach Intell 28(4):594–611
14. Yan Y, Sun J, Yu J, Sun J (2020) Small sample radar target recognition based on metric learning. IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, pp. 441–445. New York: IEEE
15. Wang Z, Wang J, Wang Y (2018) 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
16. Lv F, Wang Y, Ruan HL, Qin Y, Wang P (2021) Labeled sample augmentation based on deep embedding relation space for semi-supervised fault diagnosis of gearbox. Journal of Scientific Instrument 42(2):55–65
17. Hu T, Tang T, Lin R, Chen M, Han S, Wu J (2020) A simple data augmentation algorithm and a self-adaptive convolutional architecture for few-shot fault diagnosis under different working conditions. Measurement 156:107539
18. Chen C, Shen F, Yan R (2017) Enhanced least squares support vector machine-based transfer learning strategy for bearing fault diagnosis. Chinese Journal of entific Instrument 38(01):33–40
19. Wu ZH, Jiang HK, Zhao K, Li XQ (2019) An adaptive deep transfer learning method for bearing fault diagnosis. Measurement 151:107227
20. Li X, Hu Y, Li M, Zheng J (2019) Fault diagnostics between different type of components: A transfer learning approach. Applied Soft Computing 86:105950
21. Zheng HL, Wang RX, Yin JC, Li YQ, Lu HQ, Xu MQ (2020) A new intelligent fault identification method based on transfer locality preserving projection for actual diagnosis scenario of rotating machinery. Mech Syst Signal Process 135:106344
22. Zhang A, Li S, Cui Y, Yang W, Hu J (2019) Limited data rolling bearing fault diagnosis with few-shot learning. IEEE Access 7:110895–110904
23. Wang C, Sun H, Cao X (2021) Construction of the efficient attention prototypical net based on the time-frequency characterization of vibration signals under noisy small sample. Measurement 179:109412
24. Li W, Yang C, Peng Y, Du J (2022) A Pseudo-Siamese Deep Convolutional Neural Network for Spatiotemporal Satellite Image Fusion. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15:1205–1220
25. Jiang C, Xiao J, Xie Y, Tillo T, Huang K (2018) Siamese network ensemble for visual tracking. Neurocomputing 275:2892–2903
26. Yichi Z, Bryan P, Zhiyao D (2018) Siamese style convolutional neural networks for sound search by vocal imitation. IEEE/ACM Transactions on Audio, Speech, and Language Processing 27:429–441
27. Ahrabian K, Babaali B (2019) Usage of autoencoders and siamese networks for online handwritten signature verification. Neural Comput Appl 31(12):9321–9334
28. Li D, Tian YJ (2018) Survey and experimental study on metric learning methods. Neural Networks 105:447–462
29. Snell J, Swersky K, Zemel RS (2017) Prototypical networks for few-shot learning. Red Hook, NY, USA, pp. 4080-4090. New York: Curran Associates Inc
30. Vinyals O, Blundell C, Lillicrap T, Kavukcuoglu K, Wierstra D (2016) Matching networks for one shot learning. Red Hook, NY, USA, pp.3637–3645. Barcelona: ACM
31. Chicco D (2021) Siamese neural networks: An overview. Methods Mol Biol 2190:73–94

Publisher's Note