


A novel intelligent fault diagnosis method of bearing based on multi-head self-attention convolutional neural network

Hang Ren¹, Shaogang Liu¹, Bo Qiu², Hong Guo¹ and Dan Zhao¹ 

¹College of Mechatronic Engineering, Harbin Engineering University, Harbin 150001, China and ²CSSC Fire Equipment Co., Ltd, Jiangxi 332000, China

Research Article

Cite this article: Ren H, Liu S, Qiu B, Guo H, Zhao D (2024). A novel intelligent fault diagnosis method of bearing based on multi-head self-attention convolutional neural network. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* **38**, e9, 1–14. <https://doi.org/10.1017/S0890060423000197>

Received: 12 January 2023

Revised: 6 June 2023

Accepted: 8 August 2023

Keywords:

bearing fault diagnosis; convolutional neural network; deep learning; global information; multi-head self-attention mechanism

Corresponding author:

Dan Zhao;

Email: hezhaodan@outlook.com

Abstract

Deep learning (DL) has been widely used in bearing fault diagnosis. In particular, convolutional neural networks (CNNs) improve diagnosis accuracy by extracting excellent fault features. However, CNN lacks an explicit learning mechanism to distinguish between different fault characteristics in the input signal to the diagnosis results. This article presents a new end-to-end depth framework called multi-head self-attention convolution neural network (MSA-CNN) for bearing fault diagnosis. Firstly, we adopt a data pre-processing method that directly converts one-dimensional (1D) original signals into two-dimensional (2D) grayscale images, which is simple to implement and preserves the complete information of the original signal. Secondly, multi-head self-attention (MSA) is first constructed to aggregate the global information and adaptively assign weights to the input signal's features. Thirdly, the CNN with small-scale kernels extracted detailed local features. Finally, the learned high-level representations are fed into the full connect (FC) layer for fault diagnosis. The performance of the MSA-CNN is validated on different datasets. The results show that the proposed MSA-CNN can significantly improve fault diagnosis accuracy compared with the other state-of-the-art methods and has excellent noise immunity performance.

Introduction

Rotating machinery is indispensable in manufacturing, transportation, aerospace, and navigation (Lei et al., 2020; Xie et al., 2023). However, due to the harsh operating conditions, the transmission systems of these rotating machines will inevitably malfunction (Liu et al., 2018). Rolling bearings, as important rotating parts in mechanical equipment, are also one of the critical fault sources of mechanical equipment. Statistics show that about 30% of faults in rotating machinery are caused by bearings (Manikandan and Duraivelu, 2021). Studying bearing fault monitoring and diagnosis has significant economic and practical benefits. Therefore, bearing fault diagnosis is a crucial issue. The conventional fault diagnosis process is divided into two main steps: feature extraction and classification. However, extracting high-quality features is challenging, and manual feature extraction relies excessively on *a priori* knowledge (Chen et al., 2022). In addition, the generalization of features suffers from noise and variable operating conditions (Xiao et al., 2022).

With the rapid development of deep learning (DL), DL has been widely used in various fields, such as computer vision (He et al., 2016), medicine (Li et al., 2023), and fault diagnosis (Shao et al., 2019). DL has emerged as an effective way to overcome the drawback of conventional fault diagnosis methods (Zhang et al., 2018). He and He (2020) proposed a new deep hybrid signal processing method that combines discrete Fourier transform, inverse discrete Fourier transform, and self-coding. Li et al. (2019a) proposed a novel fault diagnosis algorithm based on sparsity and neighborhood-preserving deep extreme learning machines. Zhao et al. (2022b) proposed a new bearing fault diagnosis method based on joint distribution adaptive (JDA) and deep belief network (DBN) with an improved sparrow search algorithm (CWTSSA). Zhao et al. (2022a) proposed a vibration amplitude spectrum imaging feature extraction method using continuous wavelet transform and image conversion.

Convolutional neural network (CNN) have efficient feature extraction capabilities. Since the original signal used for fault diagnosis is one dimension (1D), scholars were the first to apply 1DCNN (Huang et al., 2019; Hao et al., 2020). Due to the local correlation of CNN, it is more suitable for processing two-dimensional (2D) images. Some researchers have converted the original signal into 2D images and conducted fault diagnosis based on 2DCNN. Liang et al. (2020) proposed a fault diagnosis method based on wavelet transform (WT). Lu et al. (2016) transformed the 1D vibration signal into a bi-spectrum contour map as the input of CNN. However, the above method of converting 1D signals into 2D images may

lose some features due to the pre-extraction of features. Therefore, Wen et al. (2018) directly converted the vibration signal into a grayscale image by sequentially implementing the pixels of the image from the original signal and using 2DCNN for bearing fault diagnosis. This method is more convenient and can achieve end-to-end learning. However, CNN focuses more on local features, the global features are equally important (Zhou et al., 2023). CNN lacks a mechanism to pay attention to features important to diagnostic results in input signals.

To overcome the above problems, some scholars have integrated the multi-head self-attention (MSA) mechanism into diagnostic models. The MSA mechanism is a module of the transformer (Dosovitskiy et al., 2021), aggregating the global information. Qin et al. (2023) proposed an enhanced MSA and CNN (EMSACNN) with two-stage feature extraction for shield machine geological condition prediction. Xiao et al. (2020) proposed a CNN and the MSA combined approach (CNN-MHSA) for detecting phishing sites. However, the above methods all use 1D sequences as input to MSA. Due to the influence of computational complexity, the length of the input sequence is limited to a certain extent. Hence, some scholars have added the MSA mechanism to CNN to obtain the weight information of features. Li et al. (2019b) proposed the combining use of dilated residual network (DRN) and MSA for speech emotion recognition (SER). Wang et al. (2020) added an MSA mechanism after the convolutional layer to build a bearing fault diagnosis model. However, some of the information will be lost after passing through CNN. Therefore, the MSA mechanism needs to aggregate complete global features before CNN. In addition, the MSA mechanism will significantly increase the number of model parameters, and none of the above models study the hyperparameter of the model.

To process long sequences and assign weights to complete global features, this article proposed an intelligent diagnosis model called MSA-CNN that combined the MSA mechanism and CNN. Firstly, we first convert the 1D original signal directly into a 2D grayscale image, which preserves the complete information of the original signal. Secondly, by applying the patch embedding method, 2D grayscale images are converted into sequences as inputs to the MSA mechanism, which process samples with longer sequences. MSA-CNN invests more attention resources into the focused area, thereby obtaining more valuable details for the target task while suppressing other useless information. Thirdly, the filtered features are input into the CNN of the small-scale convolutional kernel to extract detailed local features. Finally, the features extracted by CNN are input into FC for classification. The experimental results indicate that MSA-CNN has higher diagnostic accuracy than several state-of-the-art methods.

The remaining sections of this article are organized as follows.

Section “Network architecture” introduces the MSA-CNN and explains its workflow. Section “Experimental validation” verified that MSA-CNN performs better than some state-of-the-art methods with the CWRU dataset and visualizes the classification process by t-SNE. In addition, the role of the MSA layer is to adaptively assigned weights to different features. Meanwhile, the effectiveness of the MSA mechanism in MSA-CNN is verified by ablation experiments. In Section “Other properties of MSA-CNN”, design experiments verify that the model has strong noise immunity and performs well on other datasets. Section “Conclusions” summarizes the main work of this article and gives an outlook on future research plans.

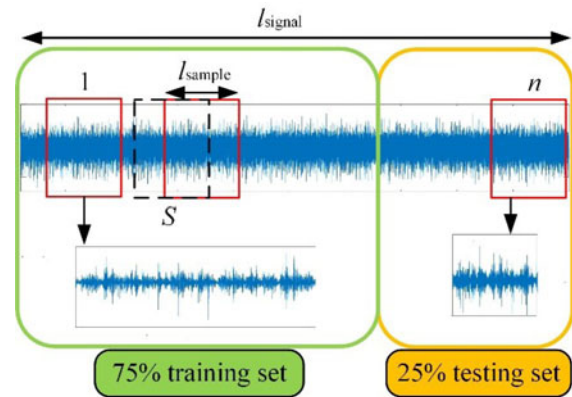


Figure 1. The data partition process.

Network architecture

Data pre-processing

Data partition

As shown in Figure 1, the vibration signals are divided into samples by “slip”. The test sets do not overlap with the training sets. The new sample is slipped by a fixed distance based on the previous sample, and the length of the slip can be calculated by the following equation:

$$S = \frac{l_{\text{signal}} - l_{\text{sample}} + 1}{n}, \quad (1)$$

where S represents the slip distance; l_{signal} represents the length of the original vibration signals; l_{sample} represents the number of data points contained in a sample; n represents the number of samples. 75% of the samples are used as the training set, and the remaining 25% are used as the test set.

Taking the bearing dataset of Case Western Reserve University (CWRU) as an example, the dataset includes three fault types for one normal operating condition. Each fault type includes three different levels of faults and a total of ten types of bearing states at loads of 0 to 3 hp. The dataset data types are:

NO: normal condition; BF_18: damage to 18 mm ball failure; BF_36: damage to 36 mm ball failure; BF_54: damage to 54 mm ball failure; IF_18: damage to 18 mm inner ring failure; IF_36: damage to 36 mm inner ring failure; IF_54: damage to 54 mm inner ring failure; OF_18: damage to 18 mm outer ring failure; OF_36: damage to 36 mm outer ring failure; OF_54: damage to 54 mm outer ring failure (Table 1).

Converting the 1D signals to 2D image

As shown in Figure 2, each sample consisting of 4096 data points can be directly converted into a 64×64 grayscale image. Each data point of the time domain samples corresponds to each pixel of the 2D grayscale image. The pixel points on the grayscale image take values in the range $[0, 255]$. The pixel values of the 1D vibration signals converted to 2D grayscale images are calculated as follows:

$$P(c, r) = \text{round}\left(\frac{A(k) - \text{Min}(A)}{\text{Max}(A) - \text{Min}(A)} \times 255\right), k = 0, 1, 2, \dots, N^2 - 1, \quad (2)$$

where $A(k)$ represents the amplitude of each vibration signal, $\text{Max}(A)$ and $\text{Min}(A)$ represent the maximum and minimum values of

Table 1. Description of bearing datasets

Label	Condition	Training set	Test set
0	BF_18	300	100
1	BF_36	300	100
2	BF_54	300	100
3	IF_18	300	100
4	IF_36	300	100
5	IF_54	300	100
6	OF_18	300	100
7	OF_36	300	100
8	OF_54	300	100
9	NO	300	100

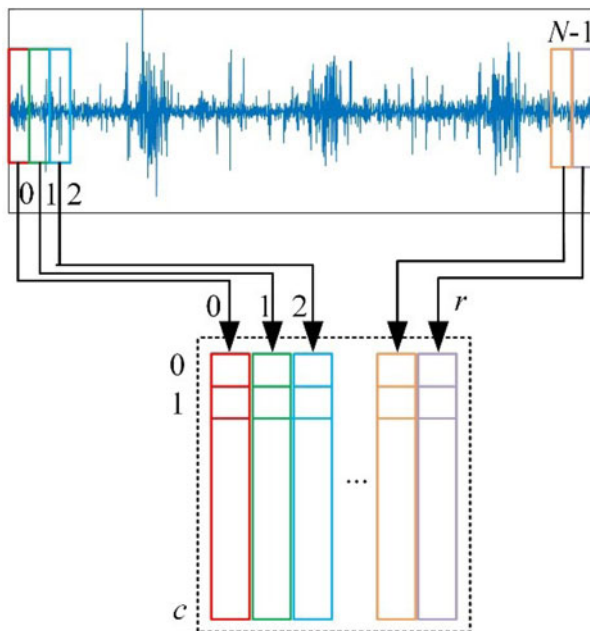


Figure 2. The original signal is converted to 2D image.

the signal, and $P(c, r)$ represents the magnitude of the pixel values in the corresponding rows and columns.

Normalization

In order to reduce the influence of data scale on the diagnostic results, the original data are first processed globally using the normalization method. The formula is given as follows:

$$x^* = \frac{(x_{max}^* - x_{min}^*) \times (x - x_{min})}{(x_{max} - x_{min})} + x_{min}^* \tag{3}$$

where $[x_{max}, x_{min}]$ represents the maximum and minimum values of each input sample; $[x_{max}^*, x_{min}^*]$ represents the normalized interval, which is taken as $[-1, 1]$. Ten types of grayscale images are shown in Figure 3.

Patch embedding

According to Vision Transformer (Dosovitskiy et al., 2021), we reshape the image $x \in \mathbf{R}^{H \times W \times C}$ into a sequence of flattened 2D

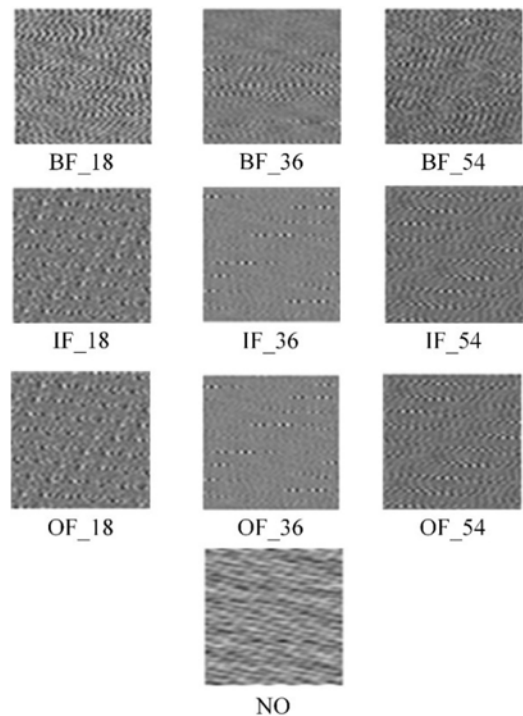


Figure 3. Grayscale images for different health status.

patches $x_p \in \mathbf{R}^{N \times (P \times P \times C)}$. We flattened the patches and mapped them to d_k dimensions with a trainable linear projection. The output of this projection is patch embedding.

The process of patch embedding is shown in Figure 4, where (H, W) is the resolution of the original image, $H = 64, W = 64$; C is the number of channels, $C = 1$; (P, P) is the resolution of each patch, $P = 16$; $N = HW/P^2$ is the resulting number of patches, which also serves as the effective input sequence length for the MSA, $N = 16$; d_k is the length of the sequence of flattened 2D patches, $d_k = 256$.

Model architecture

As shown in Figure 5, the architecture of the MSA-CNN consists of encoders and decoders.

Encoder: The encoder consists of two main sublayers, the first one is the MSA, and the second one is the multi-layer perceptrons (MLP). Each layer is connected using residuals, and layer normal precedes each sublayer.

Decoder: The decoder consists of a CNN with small-scale convolutional kernels and an MLP. The reason for using small-scale

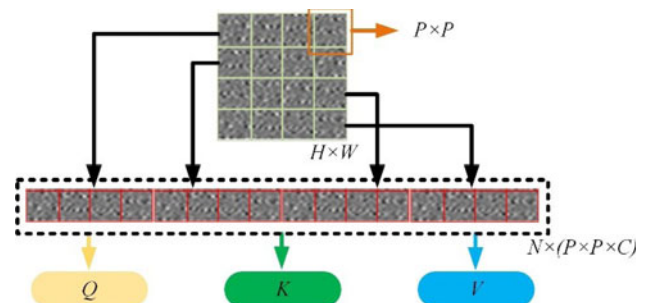


Figure 4. Patch embedding.

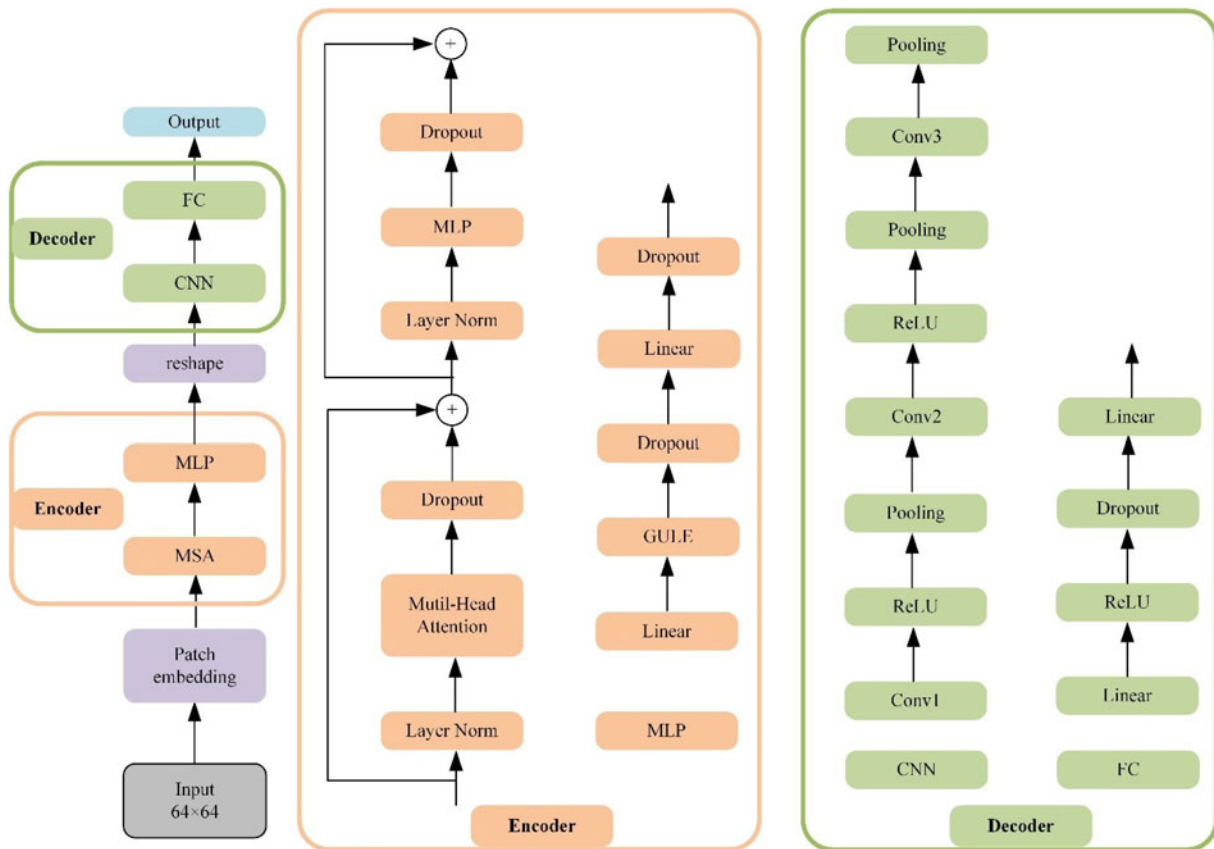


Figure 5. The architecture of the MSA-CNN.

convolution kernels is to extract detailed local features and reduce model parameters. The decoder comprises MLP and SoftMax for diagnosing fault results.

(1) Dot-product attention mechanism

As shown in Figure 6 (left), a self-attention mechanism is introduced in the first two layers of the network to improve the diagnostic accuracy. A scale dot-product attention function

mainly consists of query and key-value pairs. All the above three vectors are mapped from the same input. The similarity between query and key is calculated through the dot product to assign weights to values. According to the similarity, MSA-CNN can reinforce the learning of focused features. The following equation can compute dot-product attention.

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

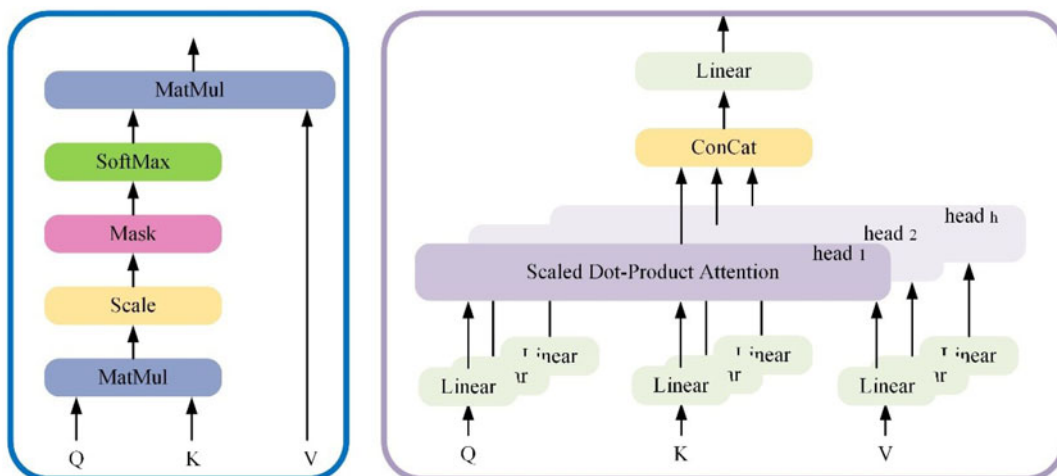


Figure 6. (Left) Scaled dot-product attention. (Right) Multi-head attention consists of several attention layers running in parallel.

where \mathbf{Q} represents the queries matrix; \mathbf{K} and \mathbf{V} represent the key matrix and the value matrix, respectively. d_k is the length of the sequence of flattened 2D patches to accelerate the convergence of the model.

(2) MSA mechanism

The MSA mechanism is shown in Figure 6 (right). It consists of a series of “scaled dot-product attention” stitched together. h heads represent using h different linear projections to extract diverse features. The MSA mechanism can be calculated by the following equation:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{concat}(\text{head}_1, \dots, \text{head}_h)\mathbf{W}^O, \quad (5)$$

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V), \quad (6)$$

where \mathbf{W}_i^Q , \mathbf{W}_i^K , \mathbf{W}_i^V are all matrices whose parameters can be learned.

(3) Convolutional layer

The convolution operation is essentially the dot product between the filter and the local region of the input data. It convolves local regions of the input signals in filter kernels, with each kernel convolving on the input vector to produce a feature vector. CNN holds the characteristic of weight sharing that can significantly reduce the number of training parameters.

$$\mathbf{x}^l(c) = \mathbf{w}_{j,k}^l \sum_{j=1}^h \sum_{k=1}^w \mathbf{x}_{j,k}^{l-1}(c) + \mathbf{b}^l, \quad (7)$$

where $\mathbf{x}^l(c)$ represents the features extracted by the l th convolutional layer; \mathbf{w}^l and \mathbf{b}^l represent the weights of the convolutional kernel and the bias terms; $\mathbf{x}_{j,k}^{l-1}(c)$ represents the 2D grayscale image; h, w represents the dimensions of the rows and columns of the 2D grayscale image. The activation function is generally added after the convolutional layer to introduce the nonlinear transformation. ReLU is the commonly used activation function calculated by the following equation:

$$\mathbf{x}^l(a) = \max\{0, (\mathbf{x}^{l-1}(a))\}, \quad (8)$$

where $\mathbf{x}^l(a)$ represents the output of the features after the activation layer; $\mathbf{x}^{l-1}(a)$ represents the features of the input before the activation layer.

(4) Pooling layer

The max-pooling layer can reduce the size of the feature map after convolution, which can ensure a significant reduction of data redundancy without affecting the valid information of the data. In addition, the pooling layer can improve the robustness of the model and it can be calculated by the following equation:

$$\mathbf{x}^l(p) = \text{down}(\mathbf{x}^{l-1}(p), s), \quad (9)$$

where $\text{down}()$ represents the down-sampling function of maximum pooling; $\mathbf{x}^l(p)$ represents the output features after maximum

pooling; $\mathbf{x}^{l-1}(p)$ represents the output of the features from the previous layer of the pooling layer; and s represents the size of the pooling.

(5) FC and dropout

After extracting features by the convolutional layer, the obtained features are spread into a 1D vector and used as input to the classifier. The FC layer can be calculated by the following equation:

$$\mathbf{x}^l(fc) = (\mathbf{w}^l)^T \mathbf{x}^{l-1}(fc) + \mathbf{b}^l, \quad (10)$$

where $\mathbf{x}^l(fc)$ represents the output of the features after the FC layer; $\mathbf{x}^{l-1}(fc)$ represents the features input before the FC layer; \mathbf{w}^l and \mathbf{b}^l represent the weight vector and the bias vector of the FC layer. To improve the generalizability of the model, a dropout strategy is used. A part of neurons will be temporarily discarded according to the set scale.

Training of the MSA-CNN model

(1) Cross-entropy loss

This section describes the training process of the proposed MSA-CNN model. The model uses the cross-entropy loss to evaluate the difference between the predicted probability distribution and the actual probability distribution of the output of the SoftMax layer, which can be calculated by the following equation:

$$L_{\text{MSA-CNN}} = H(p(x), q(x)) = - \sum_x p(x) \log q(x), \quad (11)$$

where $L_{\text{MSA-CNN}}$ represents the cross-entropy loss; $p(x)$ represents the actual probability distribution; and $q(x)$ represents the predicted probability distribution.

(2) Backpropagation

The gradient g of the output layer can be calculated by the following equation:

$$g = \frac{\partial L_{\text{MSA-CNN}}(\mathbf{w}, \mathbf{b})}{\partial \mathbf{x}}, \quad (12)$$

where g represents the gradient of the output layer. The chain rule is used to calculate the gradient of each layer, and the parameters of MSA-CNN are updated by backpropagation.

Network hyperparameter configuration

The hyperparameters of the model were selected by manual experience as follows:

- (1) The learning rate is set to 0.001.
- (2) The batch size is 32, and the epoch is set to 60.
- (3) The GELU activation function was used in the MSA mechanism. The number of heads of MSA is four, and the MLP ratio (The ratio of the number of neurons in the hidden layer to the input layer in an MLP of the MSA mechanism) is 1.
- (4) The ReLU activation function is used in the FC and convolutional layers. The dropout rate is set to 0.2.

- (5) The cross-entropy loss function was used to calculate the gradient. The stochastic gradient descent (SGD) is adopted in the optimizer, and the weight decay is set to 0.01.

The process of training

This section introduces the training process shown in Figure 7, where the whole process network parameters are updated according to the gradient backpropagation. The specific training process is listed as follows:

- (1) The 1D original signals are converted into the 2D grayscale images. These 2D grayscale images are embedded into sequences and fed into the encoder.
- (2) The global information is aggregated after an MSA module. The encoder can adaptively score the basic input features and assign weights to different features.
- (3) The output of the encoder is reshaped into a matrix as the input of the 2DCNN (Conv1–Conv2–Conv3), and small-scale convolutional kernels are used in the CNN to extract locally refined features.
- (4) The features extracted by the CNN are spread into 1D vectors for input to the FC layers (FC1–FC2) and classified by SoftMax.

The parameter value of the MSA in MSA-CNN is set according to Vision Transformer (Dosovitskiy et al., 2021). The parameter values of CNN in MSA-CNN are based on the models of LeNet-5 and Wen (Wen et al., 2018). Small-scale convolution kernels have fewer parameters and are more efficient. The specific configuration of the parameters of each layer of the network is shown in Table 2.

Experimental validation

Experimental description

The CWRU dataset (Smith and Randall, 2015) is rich in fault types and is used by many diagnostic methods. The proposed MSA-CNN model is trained and validated using the CWRU dataset at a 12 kHz sampling frequency on the motor drive side.

As shown in Table 3, the dataset is divided according to the data pre-processing method in the section “Data pre-processing”. There are 300 training set samples and 100 test set samples for each fault type. The CWRU bearing test bench is shown in Figure 8.

In order to verify the performance of the proposed model under different loads, five datasets were established under different loads from 0 to 3 as shown in Table 3. Ten types of the original signals are shown in Figure 9.

The training and test of our model are run in the Pytorch1.11 environment built on PyCharm community v2020 of the Windows 10 × 64 Professional. The experiment platform is AMD Ryzen 7 5700G CPU, 1 T hard drive, 32 G memory, NVIDIA RTX 2060 GPU, whose memory is 12 GB.

Comparison with different models

The loss curve and accuracy curve of MSA-CNN are respectively shown in Figure 10.

The advanced models for bearing fault diagnosis, i.e., 1DCNN (Abdeljaber et al., 2017), WPECNN (Ding and He, 2017), multi-head CNN (Wang et al., 2020), TSFFCNN-PSO-SVM (Xue et al., 2021), and DRSN-CW (Zhao et al., 2020) were compared with

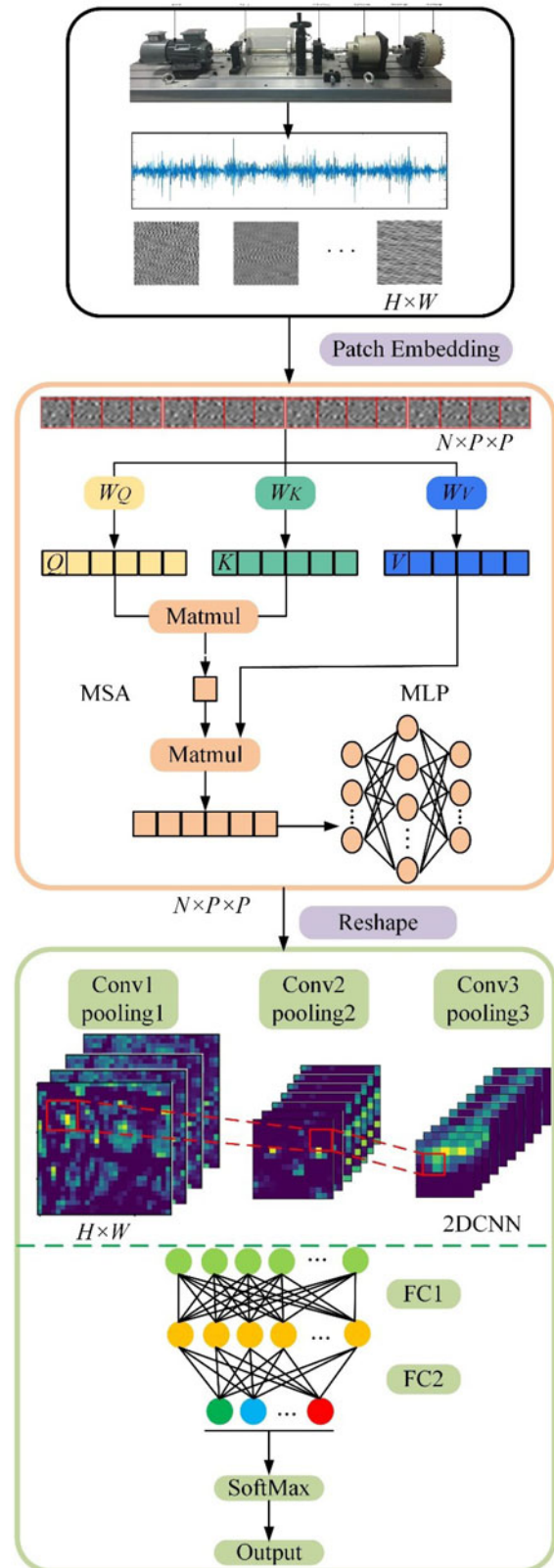


Figure 7. The training process for the MSA-CNN.

MSA-CNN to verify that the performance of MSA-CNN is superior. In this experiment, the relevant description of the above latest model is as follows.

Table 2. Parameter configuration of the MSA-CNN

Layer	Input	Output	Heads	MLP ratio
MSA	16 × 256	16 × 256	4	1
Layer	Input	Output	Kernel size	Stride
Conv1	1@64 × 64	16@30 × 30	5	2
pooling1	16@30 × 30	16@15 × 15	2	2
Conv 2	16@15 × 15	64@13 × 13	3	1
pooling2	64@13 × 13	64@6 × 6	2	2
Conv 3	64@6 × 6	256@4 × 4	3	1
pooling3	256@4 × 4	256@2 × 2	2	2
FC1	1024	512		
FC2	512	10		

Table 3. Datasets of different loads

Dataset	Load (hp)	Speed (rpm)
Dataset A	0	1797
Dataset B	1	1772
Dataset C	2	1750
Dataset D	3	1730
Dataset E	0–3	1730–1797

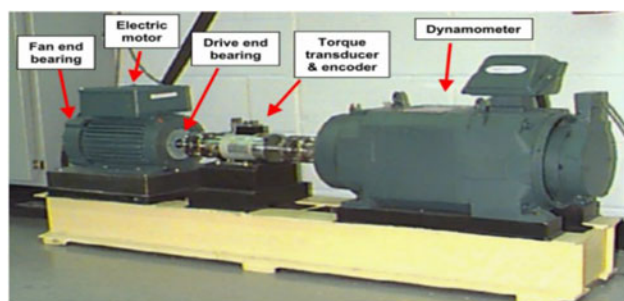


Figure 8. CWRU bearing test rig.

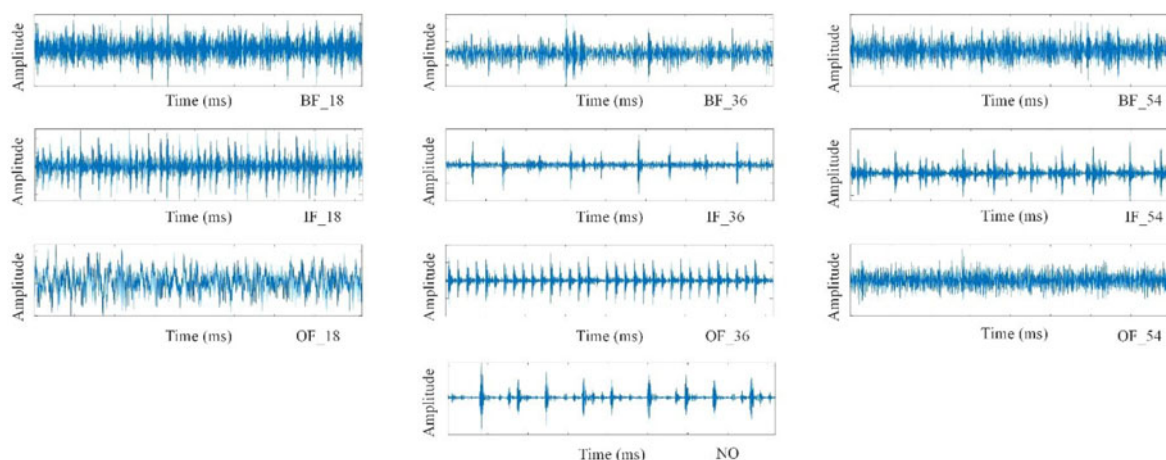


Figure 9. Ten types of the original signals.

- (1) 1DCNN: A traditional CNN consists of three convolutional layers.
- (2) WPECNN: A multi-scale feature learning method combining WPE image and deep CNN energy fluctuation for bearing fault diagnosis.
- (3) TSFFCNN-PSO-SVM: The model consists of two CNN branches. 1DCNN and 2DCNN parallel computation extract depth features, respectively, and fuse the two features by stitching to obtain more reliable diagnostic effects. In addition, Particle Swarm Optimized-Support Vector Machine (PSO-SVM) is used as the classification layer.
- (4) Multi-head CNN: A bearing fault diagnosis model that places the MSA mechanism behind the CNN to aggregate the features extracted by the CNN.
- (5) DRSN-CW: A deep residual shrinkage network combined with channel-wise thresholds (DRSN-CW). The input features are selected by the residual shrinkage building blocks, which significantly improves the anti-interference ability of the model.

The performance of MSA-CNN and other models on Datasets A–E are shown in Table 4. 1DCNN has the lowest accuracy on Datasets A–E, which proves that the diagnostic accuracy of merely CNN is limited. Compared to 1DCNN, WPECNN combines WPE diagrams and further deepens the network structure. As a result, diagnostic accuracy of WPECNN has improved. Multi-head CNN employs MSA mechanisms to improve diagnostic accuracy. DRSN-CW employ channel-wise mechanisms, and the accuracy of DRSN-CW on Datasets A–D reaches 100%. However, none of the previously mentioned networks assigns weights to global information before the CNN extracts feature. MSA-CNN has an explicit learning mechanism to distinguish the difference between different fault characteristics in the input signal and has the highest accuracy. Most importantly, the proposed MSA-CNN has higher accuracy even under complex operating conditions. This is because the MSA-CNN can consider the global information by the MSA mechanism before CNN.

The t-Distributed Stochastic Neighbour Embedding (t-SNE) is used to visualize the features of each layer. The MSA layer features, CNN3 layer features, and FC2 features are mapped into two-dimensional features. The classification visualization results of MSA-CNN on Dataset E are shown in Figure 11.

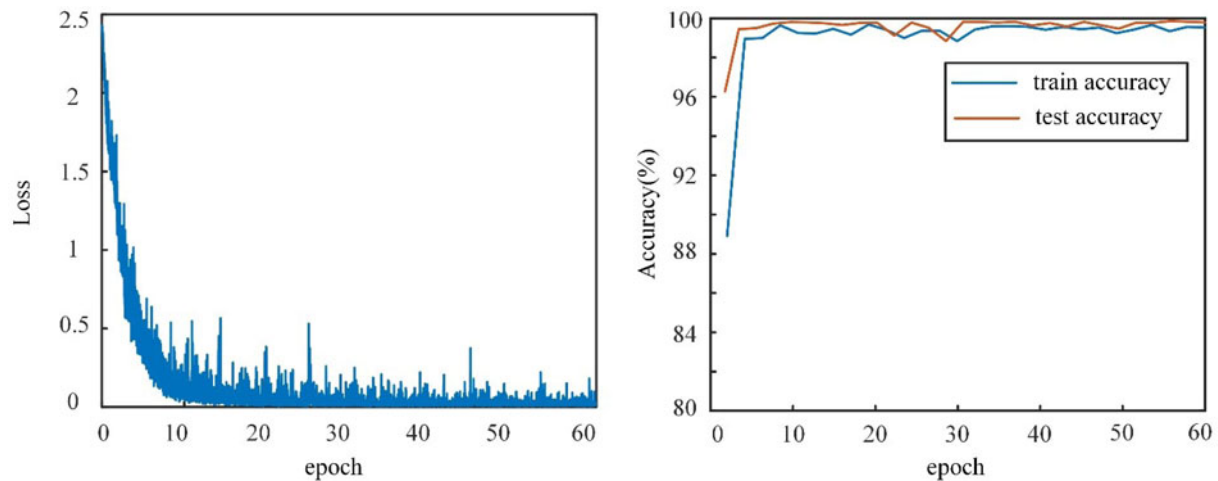


Figure 10. Loss and accuracy of MSA-CNN.

Table 4. Accuracy of different models

Model	Dataset A	Dataset B	Dataset C	Dataset D	Dataset E
1DCNN	97.6%	98.08%	98.34%	98.74%	97.2%
WPECNN	98.8%	98.8%	99.4%	99.4%	98.3%
TSFFCNN-PSO-SVM	98.5%	98.62%	98.92%	98.98%	98.23%
Multi-head CNN	99.4%	99.4%	99.8%	100%	99.2%
DRSN-CW	100%	100%	100%	100%	99.76%
MSA-CNN	100%	100%	100%	100%	99.96%

According to the results of t-SNE, there has been a tendency to aggregate the bearing data of the same type after the MSA layer. However, there is still an overlapping part of the data. After the CNN3 layer, prominent classification features were extracted,

and only a few samples were misclassified. Finally, each type is more compactly aggregated, and all fault types of bearings are precisely distinguished after two layers of MLP.

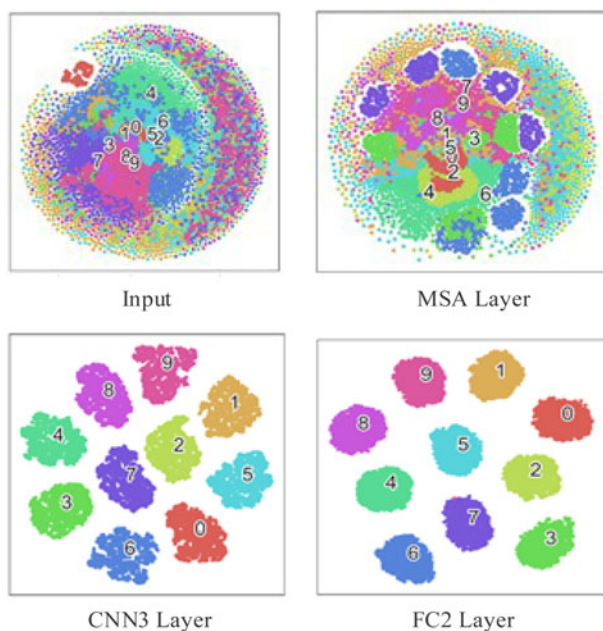


Figure 11. Visualization of MSA-CNN based on t-SNE.

Analysis of the attention layer based on the envelope spectrum

To further verify the critical role of the MSA mechanism in MSA-CNN. The envelope spectrum of the original signals and the features after the MSA layer (Attention signals) are shown in Figure 12.

Figure 12 shows no significant difference in the envelope spectrum of the four original signals within 0–2 kHz. However, the original signal envelope spectrum of the four types of bearings is significantly different after the original signal passes through MSA. Specifically, there is no significant peak in the envelope spectrum of the NO-bearing signal. However, the three faulty bearings have different degrees of peak value in the 0–0.5 kHz frequency range. In addition, bearings with faulty outer rings have the highest envelope spectral amplitude. This phenomenon shows that the MSA layer can adaptively assign weights to the original features. Therefore, after the MSA layer, the fault features in the feature space mapped from the original vibration signal are effectively utilized, resulting in better performance.

Ablation experiments

To further validate the effect of the MSA mechanism on the diagnosis results, 2DCNNs with different parameters were designed according to the reference (Wang *et al.*, 2020) and trained and tested on the CWRU dataset. CNN-A to CNN-C used the

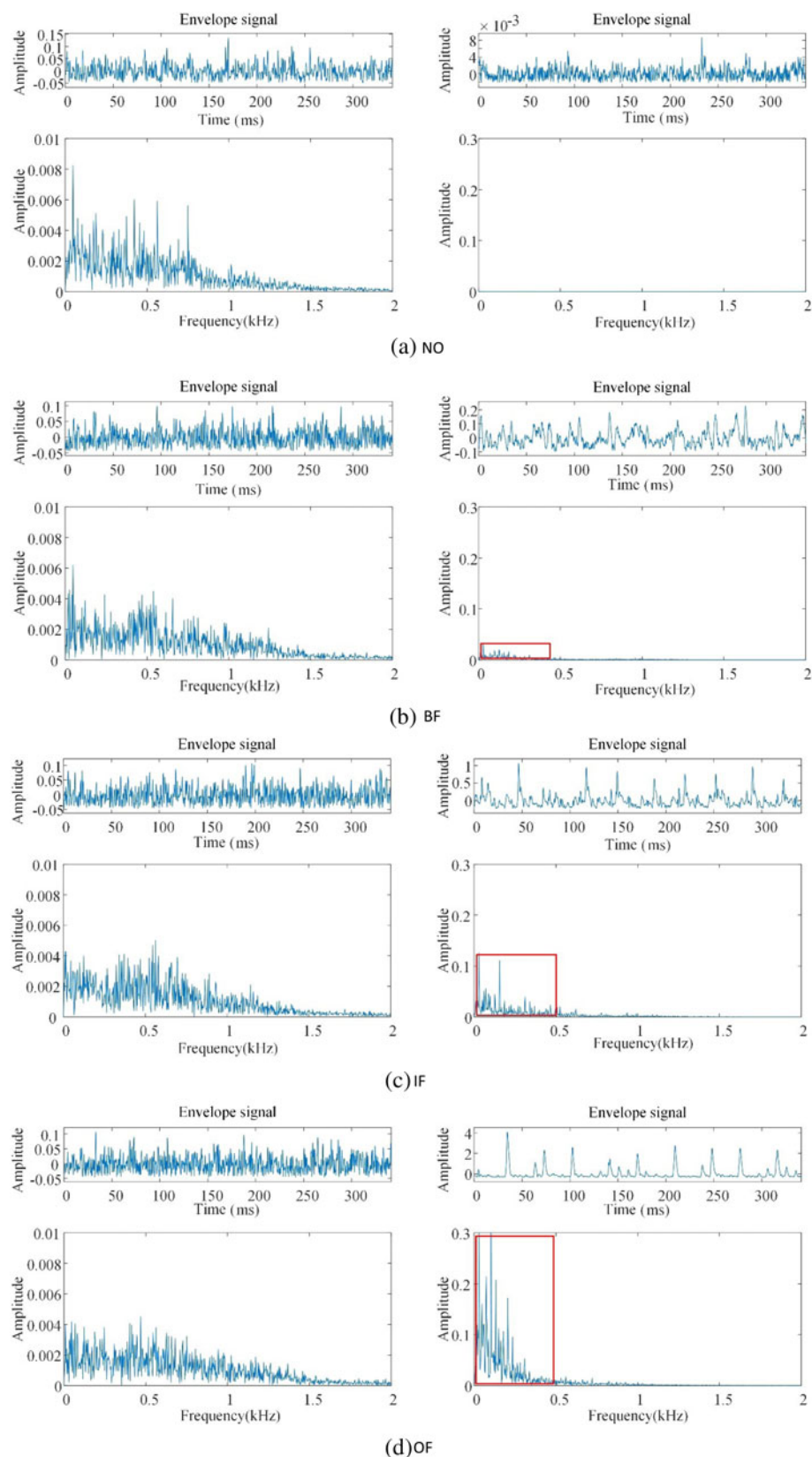


Figure 12. Envelope spectrum of the original signals and the signals aggregated by the attention layer.

traditional CNN architecture, while for CNN-D and CNN-E, the MSA mechanism was added after the CNN for feature selection. The parameters of each model are shown in Table 5, and the diagnostic results are shown in Table 6.

As shown in Table 6, the diagnostic accuracies of CNN-A, CNN-B, and CNN-C are lower than CNN-D and CNN-E on Datasets A–D. Furthermore, MSA-CNN achieves the highest accuracy compared with the other models because the

Table 5. Parameter configuration of 2DCNN

Layer	CNN-A	CNN-B	CNN-C	CNN-D	CNN-E
Input	Image size (1@32 × 32)				
Conv1	2DConv (32@3 × 3)				
pooling1	2DMaxpooling (2 × 2)				
Conv2	2DConv (64@3 × 3)				
pooling2	2DMaxpooling (2 × 2)				
Conv3	2DConv (128@3 × 3)		2DConv (192@3 × 3)		
pooling3	2DMaxpooling (2 × 2)				
Conv4	2DConv (256@3 × 3)				
pooling4	2DMaxpooling (2 × 2)				
Attention	Multi-head attention				
FC1	1024	1024	256	1024	256
FC2		256		256	
Output	10				

Table 6. Accuracy of different 2DCNN and MSA-CNN

Models	Dataset				
	A	B	C	D	E
CNN-A	98.7%	98.8%	99.9%	99.9%	83.5%
CNN-B	98.7%	98.9%	100%	100%	90.7%
CNN-C	98.8%	98.8%	99.9%	100%	94.6%
CNN-D	99.3%	98.8%	99.9%	100%	89.5%
CNN-E	99.4%	99.4%	99.8%	100%	95.1%
MSA-CNN	100%	100%	100%	100%	99.98%

MSA-CNN introduces the MSA mechanism to select the global features of the input signals adaptively. Even though CNN-D and CNN-E also introduce the MSA mechanism to select features, the accuracies of these two models are lower than that of MSA-CNN. This is because the max-pooling operation in CNN may lose some information. It is essential to add an MSA mechanism before CNN. Therefore, the proposed MSA-CNN has higher diagnostic accuracy. It is noteworthy that MSA-CNN achieves 99.96% diagnostic accuracy on the mixed-load Dataset E. The accuracy is significantly improved compared to the other five models. It is demonstrated that MSA-CNN has a strong generalization capability under complex operating conditions.

Adding an MSA layer before the CNN results in more parameters compared to a single CNN. To investigate the effect of the above parameters on the diagnostic results of the model, MSA-CNNs with different parameters are designed, and their performance on Dataset E is shown in Table 7.

As can be seen from Table 7, MSA-CNN-D reaches 99.96% accuracy for the test set on Dataset E. Overall, the accuracy of all seven models designed (MSA-CNN-A~MSA-CNN-G) remained above 99.8% when the parameters of MSA-CNN change. Even though the models become more complex by adding more layers, the models do not suffer from significant overfitting and maintain a high accuracy rate. This is because dropout and

weight decay regularization methods are used to avoid overfitting. However, increasing the number of MSA layers or increasing the MLP ratio leads to more model parameters and increase the training time of the MSA-CNN. In order to compare the number of parameters of the proposed MSA-CNN with that of the 2DCNN, the number of parameters of each model in Tables 5 and 7 is calculated as shown in Figure 13.

As can be seen from Figure 13, MSA-CNN-A and MSA-CNN-B have significantly fewer trainable parameters compared

Table 7. Accuracy of MSA-CNN configured with different parameters

Model	MSA Layer	MLP ratio	FC	Test accuracy
MSA-CNN-A	1	0.5	512-128	99.82%
MSA-CNN-B	1	1	512-128	99.80%
MSA-CNN-C	1	1	1024-256	99.92%
MSA-CNN-D	1	1	1024-512	99.96%
MSA-CNN-E	2	1	1024-512	99.92%
MSA-CNN-F	1	2	1024-512	99.86%
MSA-CNN-G	2	2	1024-512	99.92%

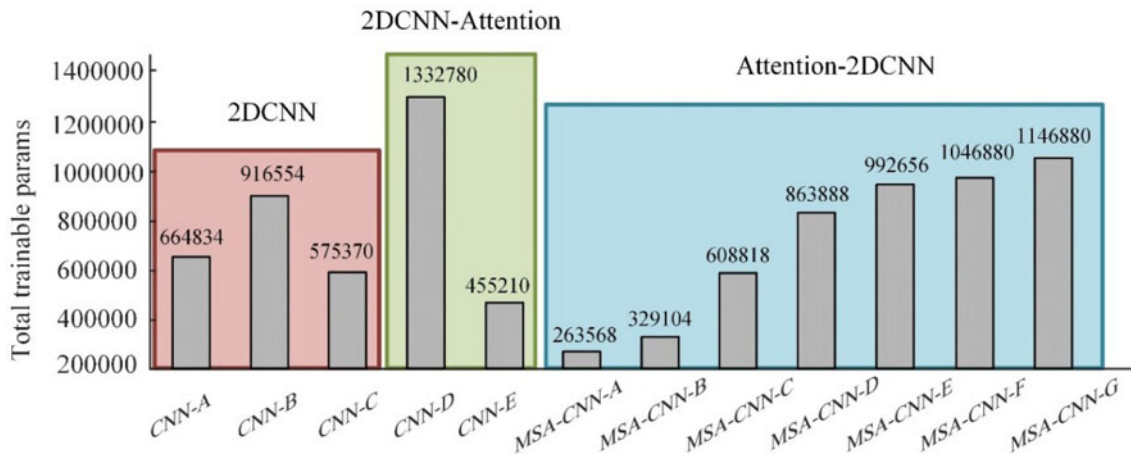


Figure 13. The number of total trainable parameters of different models.

to the other models. MSA-CNN-C does not differ much from CNN-A and CNN-C in terms of the number of parameters, but the diagnostic accuracy is improved by 16.42 and 5.32%, respectively. The remaining MSA-CNN have trainable parameters between CNN-B and CNN-D, but the diagnostic accuracy is significantly higher. In addition, compared to MSA-CNN-D, MSA-CNN-C has reduced the number of parameters by a quarter, but its accuracy has only decreased by 0.04%, which can also achieve an accuracy of 99.92%. Therefore, we believe that MSA-CNN-C is a superior model. Moreover, the original input signal length of the proposed MSA-CNN is 4096, while the input signal length of the CNN model (CNN-A~E) is 1024. The trainable parameters of the model increase accordingly when the length of the input signal of the CNN model (CNN-A~E) becomes 4096. The proposed MSA-CNN improves diagnostic accuracy using fewer trainable parameters, making the diagnostic process more efficient.

Other properties of MSA-CNN

Noise resistance study

This section investigates the noise immunity performance of MSA-CNN. It is essential to accurately diagnose faults under the influence of different noise intensities because rolling bearings have high-intensity noise in their operating environment. The criterion for evaluating the strength of signal noise is the signal-to-noise ratio (SNR). Gaussian white noise is added to the original signals on Dataset E with different SNRs to simulate the noisy environment in a natural industrial system. The SNR is defined as follows:

$$SNR = 10 \log_{10} \left(\frac{P_{\text{noise}}}{P_{\text{signal}}} \right), \quad (13)$$

where P_{noise} and P_{signal} are the power of the signals and the added Gaussian white noise, and 0 dB means that the intensity of the noise is equivalent to the original signal. On the contrary, the case of $SNR > 0$ means that the strength of the noise signal is less than the original signal. The Gaussian white noise of different intensities was added to Dataset E to restore the working environment in a realistic engineering environment. Figure 14 shows five types of signals containing noise with different SNRs.

After 30 trials on the test set, the average accuracy of the MSA-CNN for the signals containing different noise levels is shown in Figure 15. The diagnostic accuracy can reach more than 99% when the $SNR > 4$. However, the diagnostic accuracy tends to decrease with the increase in noise. The diagnostic accuracy is 94.88% when $SNR = -4$. The average diagnostic accuracy of different models containing signals with different degrees of noise is shown in Table 8.

1DCNN has the lowest accuracy, proving that a single CNN model has limited diagnostic accuracy. WPECNN uses a multi-scale approach to extract rich features and deepen the network. Therefore, the accuracy of WPECNN is slightly improved compared to 1DCNN. Due to noise interference, an explicit learning mechanism is critical to distinguish differences in different fault characteristics in the input signal. Although the multi-headed CNN uses the MSA mechanism, it uses the MSA mechanism after the CNN, so it fails to aggregate the complete global characteristics of the original signal. However, with further enhancement of noise, MSA-CNN has the highest accuracy when $SNR = -4$,

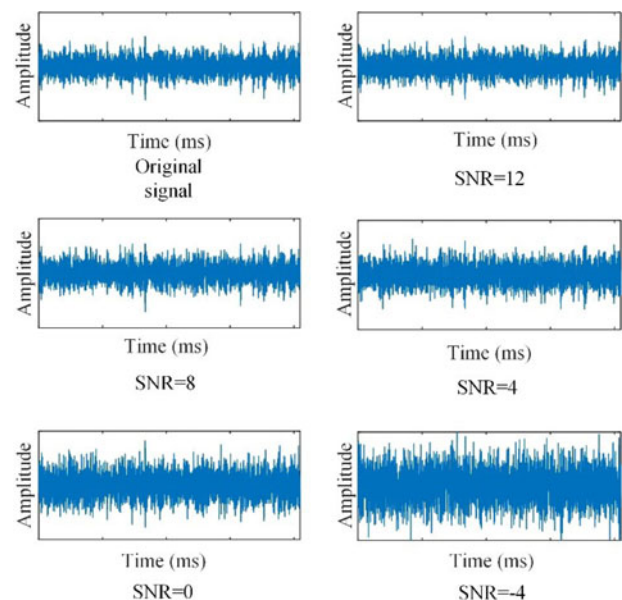


Figure 14. Signals for different SNRs.

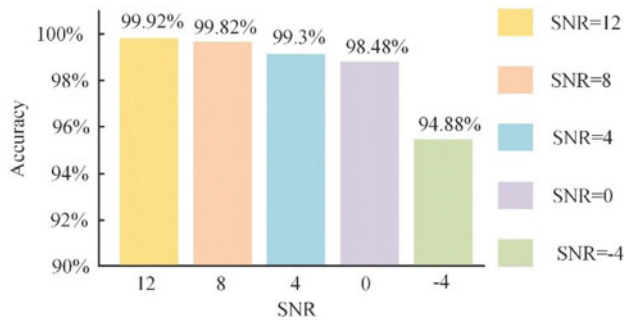


Figure 15. Accuracy of MSA-CNN on Dataset E.

Table 8. Accuracy under different degrees of noise

Model	SNR = 4	SNR = 0	SNR = -4
1DCNN	95.36%	94.25%	89.11%
WPECNN	97.11%	96.35%	90.67%
Multi-head CNN	98.41%	97.2%	92.02%
MSA-CNN	99.3%	98.48%	94.88%

demonstrating the superiority of MSA-CNN in fault diagnosis in severe noise environments.

MSA-CNN performance on other datasets

In order to verify the model performance on other datasets, the Southeast University (SU) dataset and the Jiangnan University (JU) dataset were selected for further testing.

The SU dataset conducted experiments on three datasets: induction motors, gearboxes, and bearings on Drivetrain Dynamics Simulator (Shao *et al.*, 2019). The vibration signals were collected under four operational conditions, which include two different speeds concerning two bearing loads. Bearings have four different types of faults and one healthy state for each operating condition. Therefore, this dataset has five types of data: inner ring failure, outer ring failure, rolling element failure, combined failure, and fault-free bearing. We selected 400 training samples and 100 test samples for each working condition.

Therefore, the training set contains 2000 samples, and the test set contains 500 samples.

The loss function and accuracy curves on the SU dataset are shown in Figure 16. The loss function of the MSA-CNN model tends to converge after 3 epochs, and the accuracy is 99.5%. Finally, after 60 epochs, the accuracy of the proposed MSA-CNN model is 99.86%.

The sampling frequency of the JU bearing dataset is 50 kHz. The vibration signal is obtained under three working conditions: 600, 800, and 1000 rpm. Bearings have three fault types and one healthy state for each operating condition, so this dataset has four data types (Li *et al.*, 2019b). We selected 400 training samples and 100 test samples for each working condition. Therefore, the training set contains 1600 samples, and the test set contains 400 samples.

The loss function and accuracy curves on the JU dataset are shown in Figure 17. The loss function of the MSA-CNN proposed in this article tends to converge after 20 epochs, and the accuracy is 99.5%. After 60 epochs, the proposed MSA-CNN has an accuracy of 99.68%.

Conclusions

To learn an explicit learning mechanism for distinguishing different fault features in complete input signals, this article proposes an end-to-end diagnostic model MSA-CNN for bearing fault diagnosis. The 1D original signals were directly converted into 2D grayscale images as input to MSA. The MSA layer is added before the CNN to distinguish the difference between different fault characteristics in the input signal. The experiment has verified that the proposed MSA-CNN has higher accuracy than other state-of-the-art methods. When MSA-CNN has the same number of parameters as the traditional CNN, the diagnostic accuracy of MSA-CNN is significantly improved. At the same time, the noise resistance of the model was verified. Even in severe noise environments, SNR is -4, MSA-CNN still achieved an accuracy of 94.88%, 2.86% higher than other methods. Finally, we tested the diagnostic accuracy of the SU and JU datasets. The accuracy of MSA-CNN on two datasets reached 99.96 and 99.68%, respectively.

Although this article addresses some key issues, some worth studying require further research. As a relatively large model,

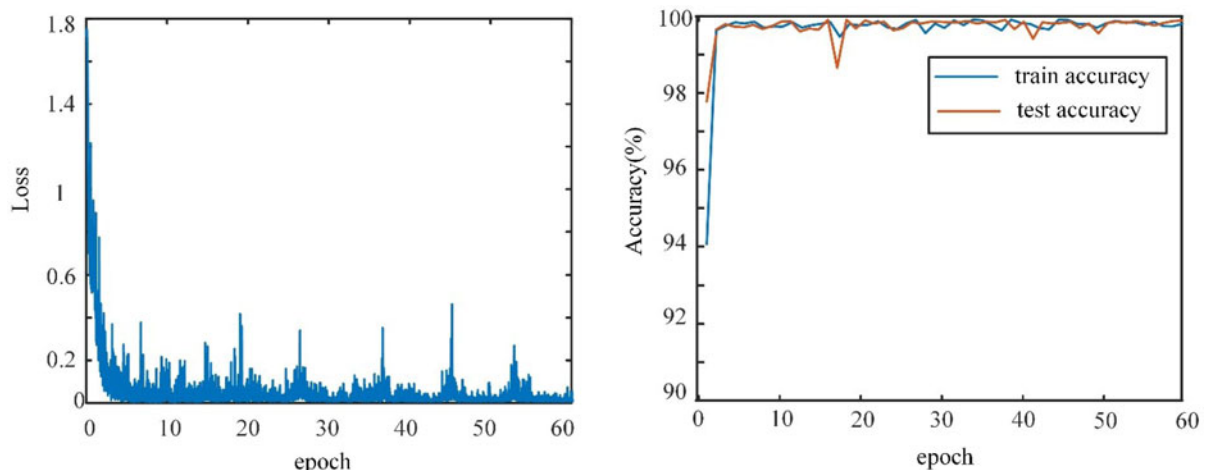


Figure 16. Loss and accuracy of MSA-CNN on the SU dataset.

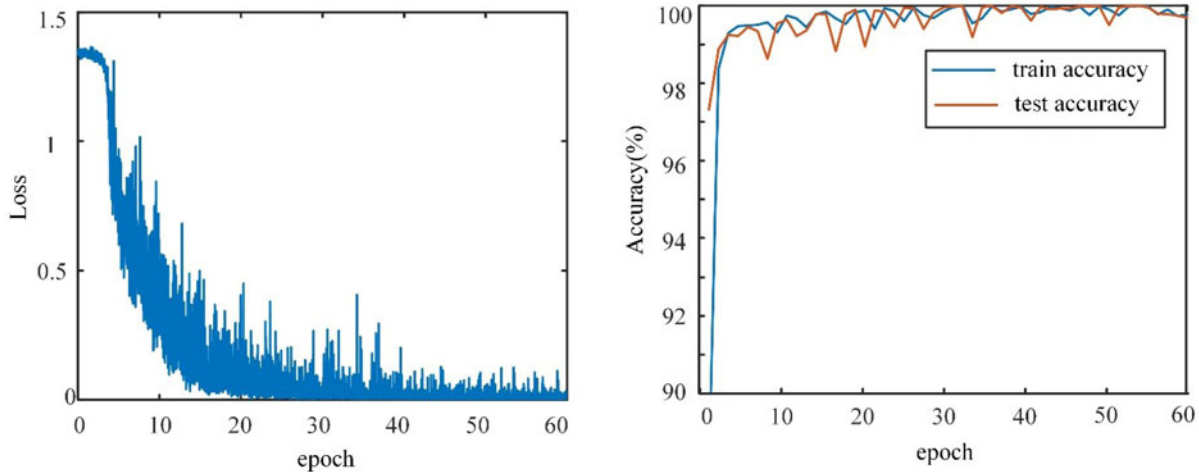


Figure 17. Loss and accuracy of MSA-CNN on the JU dataset.

MSA-CNN is prone to overfitting. Therefore, controlling the number of MS layers and using regularization methods such as dropout and weight attenuation is necessary to enhance model generalization. Finally, our future work will focus on feature weighted fusion of multi-source sensors based on the MSA mechanism for fault diagnosis.

Authors' contributions

HR: Investigation, Conceptualization, Methodology, Visualization, Writing – original draft. SL: Supervision, Writing – review and editing. BQ: prepared figures, review and editing. HG: prepared figures, review and editing. Zhao Dan: Writing – review and editing.

Funding. The authors acknowledge the funding support from the National Natural Science Foundation of China (Grant Nos. 52075111 and 51775123) and the Fundamental Research Funds for the Central Universities (Grant No.3072022JC0701). Reviewers are also appreciated for their critical comments.

Competing interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Availability of data and materials. The datasets generated during and/or analyzed during the current study are available in the CWRU dataset repository, <https://engineering.case.edu/bearingdatacenter>.

References

- Abdeljaber O, Avci O, Kiranyaz S, Gabbouj M and Inman DJ (2017) Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks. *Journal of Sound and Vibration* **388**, 154–170.
- Chen Y, Zhang D, Zhang H and Wang Q-G (2022) Dual-path mixed-domain residual threshold networks for bearing fault diagnosis. *IEEE Transactions on Industrial Electronics* **69**, 13462–13472.
- Ding X and He Q (2017) Energy-fluctuated multiscale feature learning with deep ConvNet for intelligent spindle bearing fault diagnosis. *IEEE Transactions on Instrumentation and Measurement* **66**, 1926–1935.
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houshy N (2021) An image is worth 16X16 words: transformers for image recognition at scale. *International Conference on Learning Representations*.

- Hao S, Ge F-X, Li Y and Jiang J (2020) Multisensor bearing fault diagnosis based on one-dimensional convolutional long short-term memory networks. *Measurement* **159**, 107802.
- He M and He D (2020) A new hybrid deep signal processing approach for bearing fault diagnosis using vibration signals. *Neurocomputing* **396**, 542–555.
- He K, Zhang X, Ren S and Sun J (2016) Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.
- Huang W, Cheng J, Yang Y and Guo G (2019) An improved deep convolutional neural network with multi-scale information for bearing fault diagnosis. *Neurocomputing* **359**, 77–92.
- Lei Y, Yang B, Jiang X, Jia F, Li N and Nandi AK (2020) Applications of machine learning to machine fault diagnosis: a review and roadmap. *Mechanical Systems and Signal Processing* **138**, 106587.
- Li K, Xiong M, Li F, Su L and Wu J (2019a) A novel fault diagnosis algorithm for rotating machinery based on a sparsity and neighborhood preserving deep extreme learning machine. *Neurocomputing* **350**, 261–270.
- Li R, Wu Z, Jia J, Zhao S and Meng H (2019b) Dilated residual network with multi-head self-attention for speech emotion recognition. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6675–6679.
- Li M, Zhang J, Song J, Li Z and Lu S (2023) A clinical-oriented non-severe depression diagnosis method based on cognitive behavior of emotional conflict. *IEEE Transactions on Computational Social Systems* **10**, 131–141.
- Liang P, Deng C, Wu J and Yang Z (2020) Intelligent fault diagnosis of rotating machinery via wavelet transform, generative adversarial nets and convolutional neural network. *Measurement* **159**, 107768.
- Liu R, Yang B, Zio E and Chen X (2018) Artificial intelligence for fault diagnosis of rotating machinery: a review. *Mechanical Systems and Signal Processing* **108**, 33–47.
- Lu C, Wang Y, Ragulskis M and Cheng Y (2016) Fault diagnosis for rotating machinery: a method based on image processing. *PLoS One* **11**, e0164111.
- Manikandan S and Duravelu K (2021) Fault diagnosis of various rotating equipment using machine learning approaches – a review. *Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering* **235**, 629–642.
- Qin C, Huang G, Yu H, Wu R, Tao J and Liu C (2023) Geological information prediction for shield machine using an enhanced multi-head self-attention convolution neural network with two-stage feature extraction. *Geoscience Frontiers* **14**, 101519.
- Shao S, McAleer S, Yan R and Baldi P (2019) Highly accurate machine fault diagnosis using deep transfer learning. *IEEE Transactions on Industrial Informatics* **15**, 2446–2455.
- Smith WA and Randall RB (2015) Rolling element bearing diagnostics using the Case Western Reserve University data: a benchmark study. *Mechanical Systems and Signal Processing* **64–65**, 100–131.

- Wang H, Xu J, Yan R, Sun C and Chen X** (2020) Intelligent bearing fault diagnosis using multi-head attention-based CNN. *Procedia Manufacturing* **49**, 112–118.
- Wen L, Li X and Gao L** (2018) A new convolutional neural network-based data-driven fault diagnosis method. *IEEE Transactions on Industrial Electronics* **65**, 5990–5998.
- Xiao X, Zhang D, Hu G, Jiang Y and Xia S** (2020) CNN-MHSA: a convolutional neural network and multi-head self-attention combined approach for detecting phishing websites. *Neural Networks* **125**, 303–312.
- Xiao Y, Shao H, Han S, Huo Z and Wan J** (2022) Novel joint transfer network for unsupervised bearing fault diagnosis from simulation domain to experimental domain. *IEEE/ASME Transactions on Mechatronics* **27**, 5254–5263.
- Xie C, Zhou L, Ding S, Liu R and Zheng S** (2023) Experimental and numerical investigation on self-propulsion performance of polar merchant ship in brash ice channel. *Ocean Engineering* **269**, 113424.
- Xue F, Zhang W, Xue F, Li D, Xie S and Fleischer J** (2021) A novel intelligent fault diagnosis method of rolling bearing based on two-stream feature fusion convolutional neural network. *Measurement* **176**, 109226.
- Zhang W, Li C, Peng G, Chen Y and Zhang Z** (2018) A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. *Mechanical Systems and Signal Processing* **100**, 439–453.
- Zhao M, Zhong S, Fu X, Tang B and Pecht M** (2020) Deep residual shrinkage networks for fault diagnosis. *IEEE Transactions on Industrial Informatics* **16**, 4681–4690.
- Zhao H, Liu J, Chen H, Chen J, Li Y, Xu J and Deng W** (2022a) Intelligent diagnosis using continuous wavelet transform and Gauss convolutional deep belief network. *IEEE Transactions on Reliability* **72**, 692–702.
- Zhao H, Yang X, Chen B, Chen H and Deng W** (2022b) Bearing fault diagnosis using transfer learning and optimized deep belief network. *Measurement Science and Technology* **33**, 065009.
- Zhou X, Cai X, Zhang H, Zhang Z, Jin T, Chen H and Deng W** (2023) Multi-strategy competitive-cooperative co-evolutionary algorithm and its application. *Information Sciences* **635**, 328–344.
- Hang Ren** is currently working toward the Ph.D. degree in mechanical engineering at the Institute of Shipbuilding Machinery, Harbin Engineering University, Harbin, PR China. His research interests include machinery condition monitoring, intelligent fault diagnostics, remaining useful life prediction of rotating machinery, and intelligent vibration control.
- Shaogang Liu** is currently a Professor of Mechanical Engineering at Harbin Engineering University. From 2000 to 2007, he was also the Deputy Chief Engineer of the Harbin Mechanical Research Institute of the State Forestry Administration and Chief Expert of the China Institute of Forestry Science. He is a permanent senior member of the Chinese Forest Society. His research interests include machinery condition monitoring and reliability technology, fault diagnosis, vibration control technology based on intelligent materials, and search and rescue technology based on an artificial intelligence (AI) system.
- Bo Qiu** received the M.S. degree in wood science and technology from the Chinese Academy of Forestry, Beijing, China, in 2009. From 2009 to the present, he has been researching ship fire safety and fire protection system health management technology in Jiujiang CSSC Fire Equipment Co., Ltd. His research interests include machinery condition monitoring, fire identification, and fire protection technology based on AI.
- Hong Guo** is currently working toward the Master's degree in mechanical engineering at the Institute of Shipbuilding Machinery, Harbin Engineering University, Harbin, PR China. His research interests include machinery condition monitoring, intelligent fault diagnostics, remaining useful life prediction of rotating machinery, and signal processing.
- Dan Zhao** is currently a Professor of mechanical engineering at Harbin Engineering University. She is a member of the International Society of Bionics. Her research interests include intelligent fault diagnosis, machinery condition monitoring and intelligent maintenance, signal processing, reliability design of electromechanical products, and intelligent vibration control technology.