Engineering Applications of Artificial Intelligence 126 (2023) 107031

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

# C-ECAFormer: A new lightweight fault diagnosis framework towards heavy noise and small samples

Jie Wang [a], Haidong Shao [a,*], Shen Yan [a], Bin Liu [b]

[a] College of Mechanical and Vehicle Engineering, Hunan University, Changsha, 410082, China
[b] Department of Management Science, University of Strathclyde, Glasgow, G1 1XQ, UK

ARTICLE INFO

*Keywords:*
C-ECAFormer
Heavy noise
Small samples
Fault diagnosis
Collaborative self-attention

ABSTRACT

In engineering practice, small-sample fault diagnosis of mechanical equipment towards heavy noise interference poses great challenges for the existing Transformer based intelligent models. To address these challenges, this paper proposes a new lightweight model called C-ECAFormer. Firstly, inverted residual block is used to establish signal correlations and inductive bias capability and to extract richer local feature information by varying the input channel dimensions. Secondly, ECAFormer module is designed to enhance the relationship awareness between different channels in the input signal features, thereby improving the model's attention to important channels. Finally, collaborative self-attention block is developed to facilitate spatial interaction between window local and grid global in vibration signals, reducing the number of parameters and computational complexity of the model. The results of two experiments demonstrate that the proposed approach accommodates advantages of lightweight and robustness in small-sample fault diagnosis tasks, compared to the existing mainstream Transformer and CNN fault diagnosis frameworks.

## 1. Introduction

Proper operation of mechanical equipment relies on the functioning of crucial components such as gearboxes. Malfunctions in these components can result in significant economic losses and even pose risks to personnel safety (Zhang et al., 2023a; Shao et al., 2020; Zhen et al., 2022). Due to the high-intensity usage and prolonged operation of mechanical equipment, bearings and gears often experience various types of fault modes. Therefore, employing advanced fault diagnosis methods for the detection and identification of critical components in mechanical equipment can effectively mitigate unnecessary losses (Sun et al., 2023; Wang et al., 2023; Meng et al., 2023).

Compared to traditional machine learning approaches, deep learning-based fault diagnosis techniques have reduced the reliance on prior knowledge and enable feature extraction and fault classification from complex data (Xiao et al., 2023; Hou et al., 2023; Chen et al., 2023; Zhang et al., 2023b; Shao et al., 2022; Yao and Han, 2023). Among them, intelligent fault diagnosis represented by convolutional neural network (CNN) has achieved remarkable results (Lecun et al., 1998; Sun et al., 2022; Xi et al., 2023). Jing et al. developed a gear fault diagnosis model utilizing CNN where the raw signals were transformed into spectra and used as inputs (Jing et al., 2017). Zhang et al. proposed a method called CNN with training interference that addressed the influence of noise for end-to-end bearing fault diagnosis (Zhang et al., 2018). Han et al. designed a novel framework for mechanical fault diagnosis, called deep adversarial CNN, which enhanced the generalization capability under limited training data conditions (Han et al., 2019). Hoang et al. presented a deep structure-based bearing fault diagnosis method using CNN, which exhibited high diagnostic accuracy and robustness in noisy environments (Hoang and Kang, 2019). Wu et al. introduced a novel approach for intelligent bearing fault diagnosis using a semi-supervised CNN in order to leverage unlabeled data (Wu et al., 2021). Zhao et al. combined CNN with bidirectional long short-term memory (LSTM) and proposed a deep neural network structure called convolutional bidirectional LSTM for machine health monitoring (Zhao et al., 2017).

However, CNN models primarily focus on local correlations within sequences and lack the ability to model global dependencies, making it challenging to construct relationships between long-range sequences. As a result, CNN models struggle to extract fault information from non-stationary vibration signals that are affected by heavy noise environment. Transformer determines the weights of output features by

* Corresponding author.
*E-mail address:* hdshao@hnu.edu.cn (H. Shao).

https://doi.org/10.1016/j.engappai.2023.107031
Received 25 June 2023; Received in revised form 14 August 2023; Accepted 22 August 2023
Available online 27 August 2023

calculating the similarity between this sequence and the rest of the sequences, and has strong feature extraction and long sequence modelling capabilities, which has been widely used in natural language processing and computer vision in recent years (Vaswani et al., 2017; Dosovitskiy et al., 2020). In the past two years, several researchers have applied Transformer to mechanical fault diagnosis in the presence of heavy noise interference. In 2022, Tang et al. constructed Transformer architecture for rotating machinery fault diagnosis in variable condition problems, using CWRU dataset as the dataset with 400 samples per class of training set (Tang et al., 2022). In 2022, Li et al. designed an interpretable deep learning model, variational attention-based transformer network, in order to uncover the connections within the signals, and used 1000 samples per class in the experiment, of which 70% were used as training samples (Li et al., 2022a). In 2023, Tian et al. introduced a mechanical fault diagnosis method called wavelet-based self-attention network. This method effectively suppressed noise interference in both the time and frequency domains. The training dataset consisted of 210 samples per class (Tian et al., 2023). In 2023, Zhou et al. proposed an industrial process optimization vision transformer (ViT) method to explore the global acceptance domain provided by the self-attentive mechanism of ViT in fault detection and diagnosis. For each operating condition, a training sample of 4000 was used (Zhou et al., 2023). Although the Transformer models mentioned above have powerful feature extraction capabilities, they often suffer from a large number of parameters and a complex computational method. Additionally, due to the lack of local signal correlations and spatial inductive bias, these models typically require a substantial amount of training samples to ensure effective performance. As a result, the existing fault diagnosis methods face challenges when applied to small samples.

In this paper, a novel lightweight model called C-ECAFormer has been proposed for fault diagnosis towards heavy noise and small samples. By analysing two different experimental data, the results demonstrate that the proposed method effectively reduces the reliance on samples and improves feature extraction capabilities under strong noise interference. The main contributions of this paper are as follows:

(1) Inverted residual block is used to establish local signal correlations and inductive bias capability and to extract richer feature information by varying the input channel dimensions.
(2) ECAFormer module is designed to enhance the relationship awareness between different channels in the input signal features, thereby improving the model's attention to important channels.
(3) Collaborative self-attention block is developed to facilitate spatial interaction between window local and grid global in vibration signals, reducing the number of parameters and computational complexity of the model.

The rest of the paper is organised as follows: Section 2 briefly reviews the related works. Section 3 describes the details of the proposed model and the overall framework. Section 4 demonstrates the advantages of the proposed method through two sets of comparative experiments. Section 5 concludes the paper.

## 2. Related works

### 2.1. Efficient channel attention module

The structure of the efficient channel attention (ECA) module is illustrated in Fig. 1. (Wang et al., 2020). The ECA module consists of two steps. Firstly, the spatial dimension of the input signal features is compressed by downsampling with global average pooling to obtain a representative value for each channel, thus reflecting the overall characteristics of each channel. Secondly, the features are scaled and weighted by learning the relationships between different channel dimensions using convolutional layers. More important channels will be scaled and weighted more, while less important channels will be weakened. The output values of the network are then mapped to the range of 0–1 using the Sigmoid activation function to obtain channel weights. These weights are then applied to the original signal through element-wise product. The specific formula is as follows:

$$\chi_{ECA} = Sigmoid(Conv(GAP(X))) \tag{1}$$

$$X_{out} = \chi_{ECA} \odot X \tag{2}$$

where $X \in R^{C\times N}$ is the input; $C$ denotes the input channel dimension; $N$ represents the input length; $GAP(\cdot)$ is the global average pooling; $Conv(\cdot)$ represents the convolution function; $Sigmoid(\cdot)$ represents the activation function; $\chi_{ECA}$ is the calculated channel weight; $\odot$ is the Element-wise product and $X_{out} \in R^{C\times N}$ is the output.

### 2.2. Transformer

A typical Transformer consists of two parts: the encoder and the decoder (Vaswani et al., 2017). The main components of the encoder and the decoder include a multihead self-attention module (MSA), a feed-forward neural network (FFN), a residual connection and layernorm, the structure of MSA block is shown in Fig. 2.

In the multihead self-attention module, as the first step, the input is linearly mapped to generate the query matrix, key matrix and value matrix, which is to project the input into a higher-dimensional space, enabling subsequent computations to consider more information. The query matrix, key matrix and value matrix formed in the first step are further linearly transformed separately. The purpose here is to generate the query matrix, key matrix and value matrix in each head. The specific formula is as follows:

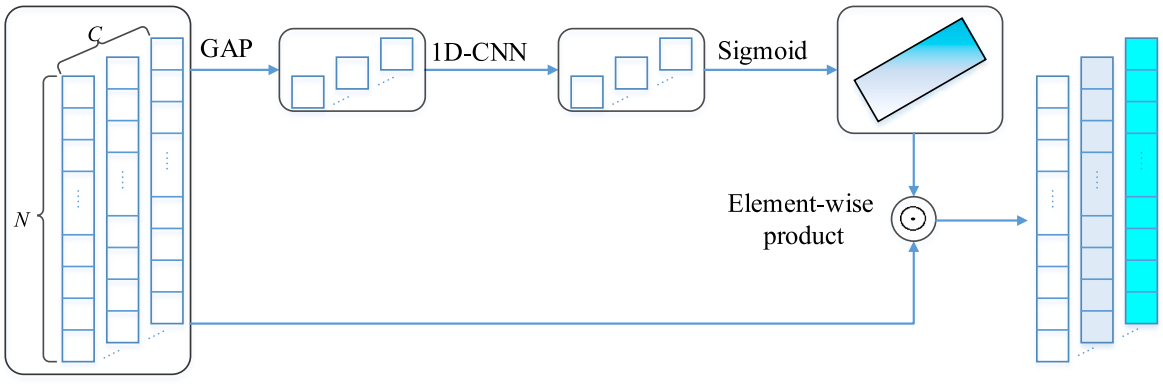


**Fig. 1.** The architecture of ECA module.

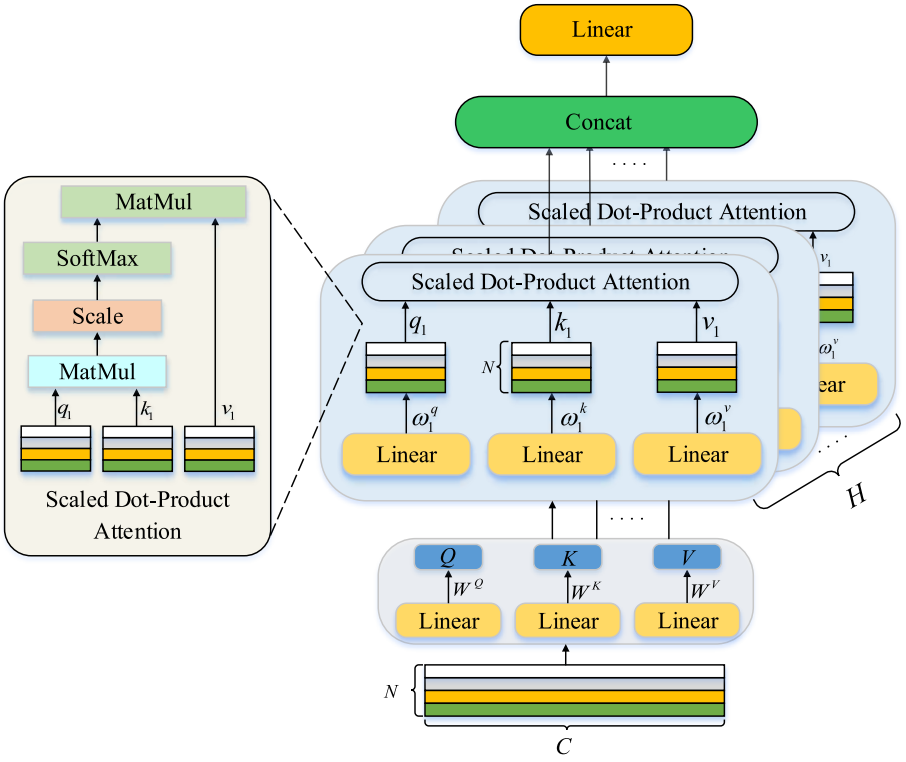


**Fig. 2.** The architecture of MSA block.

$$Q = xW^Q, K = xW^K, V = xW^V$$
$$Q_h = QW_h^Q, K_h = KW_h^K, V_h = VW_h^V \tag{3}$$

where $x \in R^{N_{MSA} \times C}$ is the input; $W^Q, W^K, W^V \in R^{C \times C}$ denote the weights of the respective linear mappings; $Q_h, K_h, V_h \in R^{N \times C_H}$ $(h = 1, 2, \cdots, H)$ represent the query matrix, key matrix and value matrix in each head, respectively; $W_h^Q, W_h^K, W_h^V \in R^{C \times C_H}$ $(h = 1, 2, \cdots, H)$ are the weights of the corresponding transformations; $H$ is the number of heads in the attention mechanism and $C_H$ is the feature dimension in each head, where $C = C_H \times H$.

Subsequently, the dot product of $Q_h$ and $K_h$ is computed. This result is then scaled and normalized to obtain the attention weights, which is further multiplied by the $V_h$ to obtain the weighted features to calculate the similarity between the input vectors. The specific formula is as follows:

$$A_h = SoftMax\left(\frac{Q_h K_h^T}{\sqrt{C_H}}\right) V_h \tag{4}$$

where $SoftMax(\cdot)$ represents SoftMax normalization operation, and $A_h \in R^{N \times C_H}$ is the result of the attention calculation in the $h$-th head.

Finally, the attention calculations in all heads are concatenated and linearly mapped to obtain the final result. The specific formula is as follows:

$$MSA = Concat(A_1, A_2, \cdots, A_H)\omega^Y \tag{5}$$

where $MSA \in R^{N \times C}$ is the output of the multihead self-attention layer; $\omega^Y \in R^{C \times C}$ is the weight of the linear mapping.

## 3. Proposed method

### 3.1. Inverted residual block

To address the limitations of Transformer in establishing local signal correlations and spatial inductive bias capability, and to enhance the model's ability to extract underlying input sequence features while preserving local information, the inverted residual block (Sandler et al., 2018) was added before the attention calculation module.

As shown in Fig. 3, to maintain local information, each channel of the input signal undergoes pointwise convolution, resulting in an up-dimensioning of the channel. Next, depthwise convolution is applied to the up-dimensioned features. Finally, pointwise convolution is utilized to reduce the dimensionality of the multi-scale features, generating new signal features. By incorporating the inverted residual block with stride of 2 for downsampling in the first CSA block in each stacked CSA block, the model can learn richer features by continuously changing the channel dimensionality. Additionally, it leverages the inherent

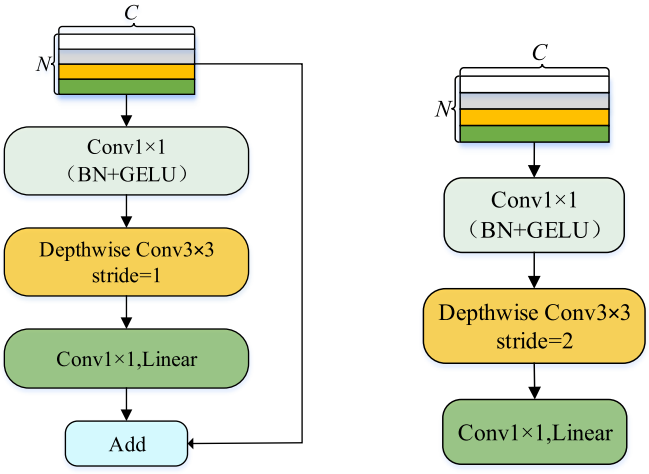


**Fig. 3.** The architecture of inverted residual block: (a) stride = 1; (b) stride = 2.

inductive bias of convolutions, thereby enhancing the model's generalization capability and trainability. The remaining CSA blocks in each stacked CSA block use stride of 1, to convey important information and gradients while reducing information loss.

### 3.2. Development of ECAFormer

In the self-attention mechanism of Transformer, the focus is primarily on computing the similarity between data features to determine the relevance of different regions in the input sequence. By employing the multihead mechanism, the model calculates the correlations between different channels. However, each head can only attend to a portion of the input channels, which may not fully capture the correlations among all channels in the input feature map. This becomes especially problematic when the number of input channels is large, as it may lead to the neglect of correlations between some channels. To address this issue, we propose the ECAFormer module, as shown in Fig. 4.

After the self-attention calculation is performed, the H-ECA module calculates the correlations between different channels within each head, so as to integrate the information of different channels in the respective heads, enhance the ability to perceive the relationship between different channels in the input signal features, and improve the model's attention to important channels. The specific formula is as follows:

$$\chi^{'}(i) = \chi_{ECA}(i) \odot A_i \tag{6}$$

$$EFr = Concat(\chi^{'}(1), \chi^{'}(2), \chi^{'}(H))\omega^{o} \tag{7}$$

where $A_i \in R^{N\times C_H}$ $(i = 1, 2, \cdots, H)$ represents the attention matrix obtained from the attention calculation in the $i$-th head; $\chi_{ECA}(i)$ is the channel weight in the $i$-th head; $\chi^{'}(i) \in R^{N\times C}$ is the matrix calculated by the channel attention mechanism in the $i$-th head; $\omega^{o} \in R^{C\times C}$ is the weight of the linear mapping and $EFr \in R^{N\times C}$ is the final output.

### 3.3. Construction of collaborative self-attention block

The typical Transformer's MSA block module uses a more complex matrix multiplication method when calculating the global feature information of the input signal, which significantly increases the computational cost. To address this issue, we develop the collaborative self-attention (CSA) block, as shown in Fig. 5. The CSA block employs two algorithms, namely, window local attention and grid global attention, to compute attention on the input signal data. This approach is able to reduce the computational complexity of traditional attention mechanisms from quadratic to linear complexity without losing non-locality.

First, in the window local attention module, the input data $\chi \in R^{N\times C}$ is unfolded into $X_b(i) \in R^{N_P\times C}$ $(i = 1, 2, \cdots, n)$ in a non-overlapping manner using a window of fixed size $N_P$, then attention calculation is performed on each window independently, and finally, the attention results from all windows are refolded into $\chi \in R^{N\times C}$ using the window fold operation. The specific formula is as follows:

$$W_i = EFr(X_b(i)) \tag{8}$$

$$Win = WF(W_1, W_2, \cdots, W_n) \tag{9}$$

where $W_i \in R^{N_p\times C}$ $(i = 1, 2, \cdots, n)$ is the result of each window calculated by the ECAFormer module; $N_P$ represents the dimension size of the window; $n$ represents the number of window, where $N = N_p \times n$; $EFr(\cdot)$ represents the operation that the input is calculated by our improved ECAFormer model; $WF(\cdot)$ denotes the window fold operation and $Win \in R^{N\times C}$ represents the final result obtained after window fold operation.

After the computation in the window local attention layer, the results are fed into a FFN layer and then passed to the grid global attention layer. In the global interactive attention layer, a non-overlapping window of size $N_G$ is used to partition the feature map $X_{in} \in R^{N\times C}$ into

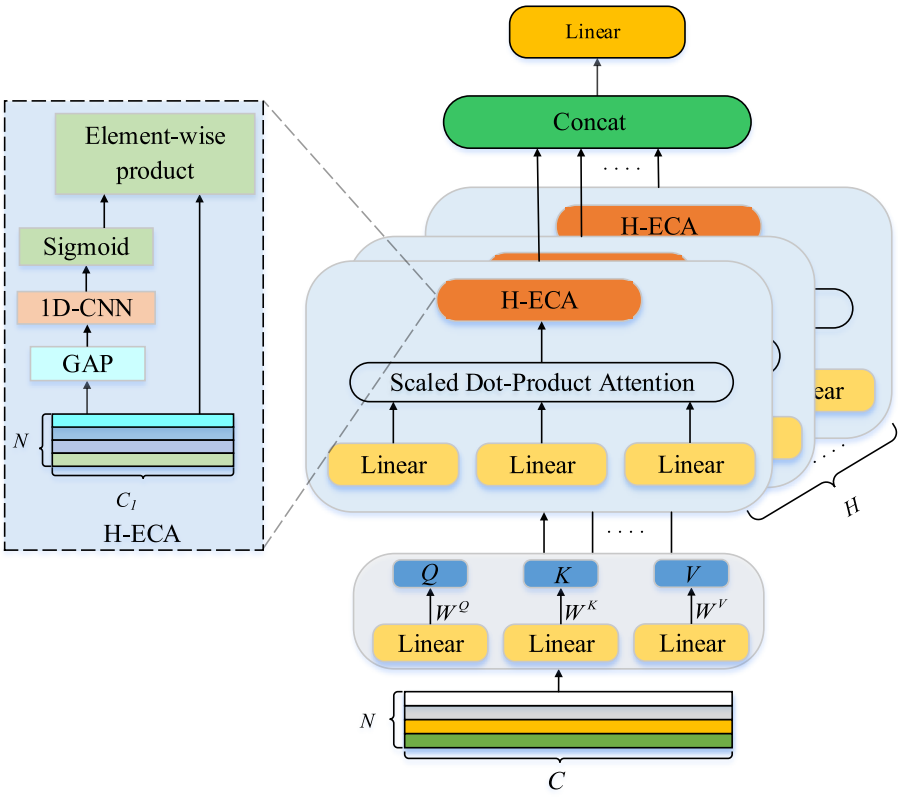


**Fig. 4.** The architecture of ECAFormer.

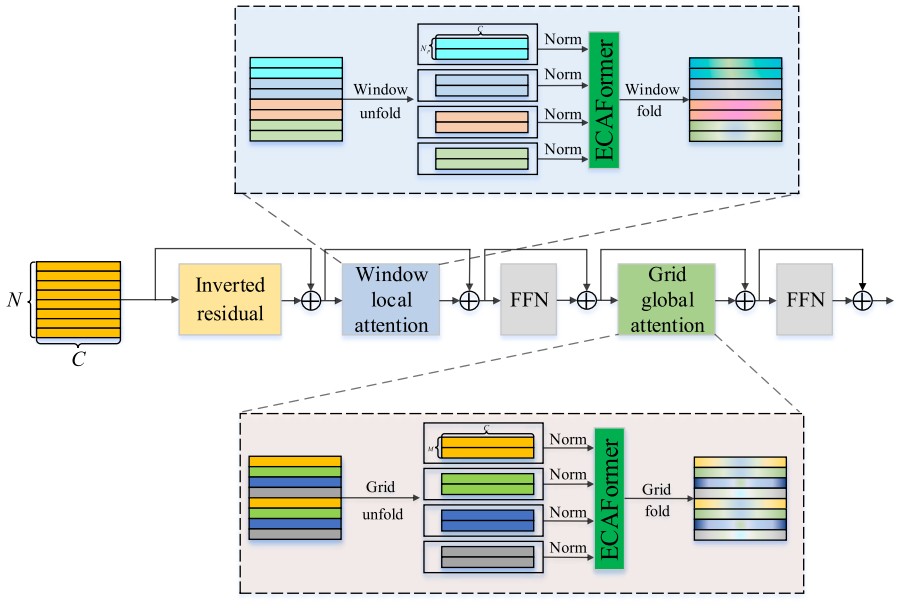


**Fig. 5.** The architecture of collaborative self-attention block.

patches, each patch with the size of $N_G \times C$, then, the tokens at the same position are extracted and refolded into patches of size $X_G(i) \in R^{M\times C}$ $(i = 1, 2, \cdots, N_G)$, followed by independent attention calculation on each patch. To preserve the spatial order of the signal data within each patch, the unfolded patches, $G_i \in R^{M\times C}$, are refolded to $X_{in} \in R^{N\times C}$ using the grid fold operation. The output is then passed through another FFN layer. The specific formula is as follows:

$$G_i = EFr(X_G(i)) \tag{10}$$

$$Grid = GF(G_1, G_2, \cdots, G_{N_G}) \tag{11}$$

where $G_i \in R^{M\times C}$ $(i = 1, 2, \cdots, N_G)$ is the result of each grid calculated by the ECAFormer module; $N_G$ represents the dimension size of the grid; $M$ represents the number of grid, where $N = N_G \times M$; $EFr(\cdot)$ represents the operation that the input is calculated by our improved ECAFormer model; $GF(\cdot)$ denotes the grid fold operation and $Grid \in R^{N\times C}$ represents the final result obtained after grid fold operation.

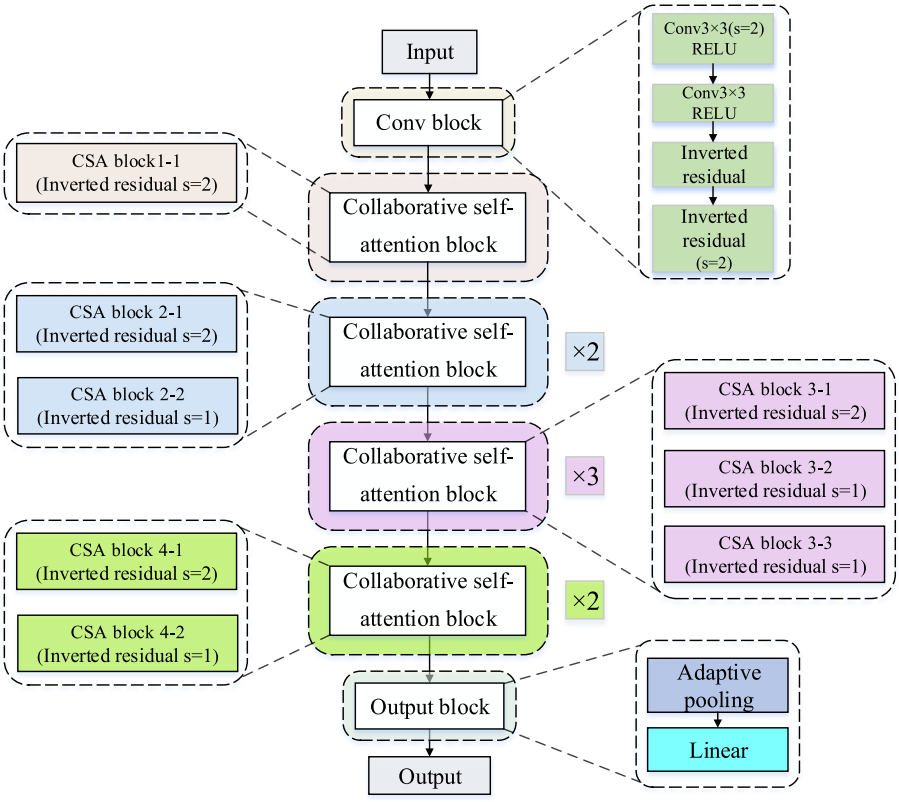


**Fig. 6.** The architecture of C-ECAFormer.

### 3.4. C-ECAFormer architecture

The structure of C-ECAFormer, as shown in Fig. 6, consists of a Conv block, four collaborative self-attention blocks, and an output block.

The Conv block contains two convolutional layers and two inverted residual blocks, which are used to reduce the dimensionality of the input signal and obtain information about the surrounding region, while increasing the number of channels of the input signal, capturing the details and features of the input signal and improving the expressiveness of the model.

The core component of the C-ECAFormer is the four collaborative self-attention blocks. In each CSA block, the first inverted residual block employs stride of 2 to change the receptive field size, while the remaining inverted residual block in the CSA blocks use stride of 1. Then, the window local attention layer is applied to compute the attention within local windows, enabling the extraction of detailed features. Subsequently, the FFN layer is utilized to the output features. Finally, grid global attention is applied to calculate the attention across global grids, establishing global dependencies and capturing long-range relationships. The FFN layer is further utilized to the output features. During the attention computation, H-ECA is incorporated to calculate the inner channel correlations within each head, enhancing the model's ability to capture the key information of the important channels.

The output block consists of an average pooling for focusing features in the temporal dimension and a linear transformation for mapping the extracted high-dimensional features into the fault diagnosis classification dimension.

Table 1 summarises the parameter and signal shape for each block in the C-ECAFormer model for an input signal length of 1024 sampling points.

**Table 1**
The parameters and signal shape of each block in the constructed C-ECAFormer.

| Blocks | Modules | Parameters | Signal Shape |
|---|---|---|---|
| Conv block | Convolution | $d = 32$; $k = 3$; $s = 2$ | 512 × 32 |
| | Convolution | $d = 32$; $k = 3$; $s = 1$ | 512 × 32 |
| | Inverted residual block | $d = 64$; $k_c = 1,3,1$; $s_c = 1,1,1$ | 512 × 64 |
| | Inverted residual block | $d = 64$; $k_c = 1,3,1$; $s_c = 1,2,1$ | 256 × 32 |
| Collaborative self-attention block 1 | CSA block 1-1 | $d = 64$; $k_c = 1,3,1$; $s_c = 1,2,1$; $h = 8$; $N_P = 16$; $M = 16$; $f = 4$ | 128 × 64 |
| Collaborative self-attention block 2 | CSA block 2-1 | $d = 96$; $k_c = 1,3,1$; $s_c = 1,2,1$; $h = 8$; $N_P = 16$; $M = 16$; $f = 4$ | 64 × 96 |
| | CSA block 2-2 | $d = 96$; $k_c = 1,3,1$; $s_c = 1,1,1$; $h = 8$; $N_P = 16$; $M = 16$; $f = 4$ | 64 × 96 |
| Collaborative self-attention block 3 | CSA block 3-1 | $d = 128$; $k_c = 1,3,1$; $s_c = 1,2,1$; $h = 8$; $N_P = 16$; $M = 16$; $f = 4$ | 32 × 128 |
| | CSA block 3-2 | $d = 128$; $k_c = 1,3,1$; $s_c = 1,1,1$; $h = 8$; $N_P = 16$; $M = 16$; $f = 4$ | 32 × 128 |
| | CSA block 3-3 | $d = 128$; $k_c = 1,3,1$; $s_c = 1,1,1$; $h = 8$; $N_P = 16$; $M = 16$; $f = 4$ | 32 × 128 |
| Collaborative self-attention block 4 | CSA block 4-1 | $d = 256$; $k_c = 1,3,1$; $s_c = 1,2,1$; $h = 8$; $N_P = 16$; $M = 16$; $f = 4$ | 16 × 256 |
| | CSA block 4-2 | $d = 256$; $k_c = 1,3,1$; $s_c = 1,1,1$; $h = 8$; $N_P = 16$; $M = 16$; $f = 4$ | 16 × 256 |
| Output block | Linear | $d = C_n$ | $C_n$ |

**Remarks:** $d$ represents the output dimension, $k$ represents the kernel size in the CNN, $s$ represents the stride in the CNN, $k_c$ represents the kernel size in the inverted residual block, $s_c$ represents the stride in the inverted residual block, $h$ represents the number of heads, $N_P$ represents the size of the window in the window local attention, $M$ represents the size of the grid in the grid global attention, $f$ represents the feedforward factor in the FFN, and $C_n$ represents the number of conditions.

### 3.5. Applications of the C-ECAFormer-based fault diagnosis method

In this paper, a novel lightweight model called C-ECAFormer has been proposed for fault diagnosis towards heavy noise and small samples. The specific procedure is shown in Fig. 7.

Step 1: Vibration signals are collected from the gearbox through sensors. The collected data is divided into training dataset and test dataset by using a sliding window.
Step 2: The training dataset is fed into a C-ECAFormer model consisting of a Conv block, some collaborative self-attention blocks and an output block for training. The Adam optimizer is used to dynamically adjust the learning rate and update the model weights. The training process selects cross-entropy loss function for backward gradient propagation.
Step 3: The test dataset is fed into the well-trained model for fault diagnosis and the diagnosis results are visualised and analysed.

## 4. Case study

### 4.1. Case 1: fault diagnosis of planetary gearbox towards heavy noise and small samples

#### 4.1.1. Data preparation of case 1

The XJTU Gearbox dataset is collected by the Institute of Aero-engine at Xi'an Jiaotong University, China (Li et al., 2022b). The experimental platform is depicted in Fig. 8(a), which consists of driving motor, controller, planetary gearbox, parallel gearbox, and brake. Two accelerometers (PCB352C04) are installed in the X and Y directions on the planetary gearbox to collect vibration signals. In the experiments, four types of gear faults and four types of bearing faults are artificially introduced to the planetary gearbox, as shown in Fig. 8(b). For simplicity, each condition is assigned a different label, as shown in Table 2. The motor speed is set to 1800 r/min, and the sampling frequency is set to 20480Hz during the experiments.

The vibration signals are divided using a sliding window, where each sample contained 1024 data points. The consecutive windows have a 30% overlap. Each data point has two channels, representing the vibration signals from the X and Y directions. To test the performance of the model under conditions of limited samples and strong noise interference, 30 training samples and 100 testing samples are selected for each healthy condition. Additionally, Gaussian white noise with a signal-to-noise ratio (SNR) of 2 is added to each sample.

$$SNR = 10 \times \log_{10}\left(\frac{P_s}{P_n}\right) \tag{12}$$

where $P_s$ indicates the effective power of the input signal, and $P_n$ represents the power of the added random noise.

The running configuration is described as follows: the software is Pytorch 1.12.1; GPU is GTX1660s; CPU is i5-10400F.

#### 4.1.2. Result analysis of case 1

To evaluate the performance of the proposed model, we select four end-to-end fault diagnosis models based on Transformer-CNN: CLFormer (Fang et al., 2021), Convformer-NSE (Han et al., 2023), MaxVit (Tu et al., 2022), and MobileVit (Mehta and Rastegari, 2021). Additionally, we include a widely used CNN fault diagnosis model, ResNet18 (He et al., 2016), as a baseline for comparison. To reduce the impact of random errors, each experiment is repeated five times, and training is conducted for 100 iterations in each experiment.

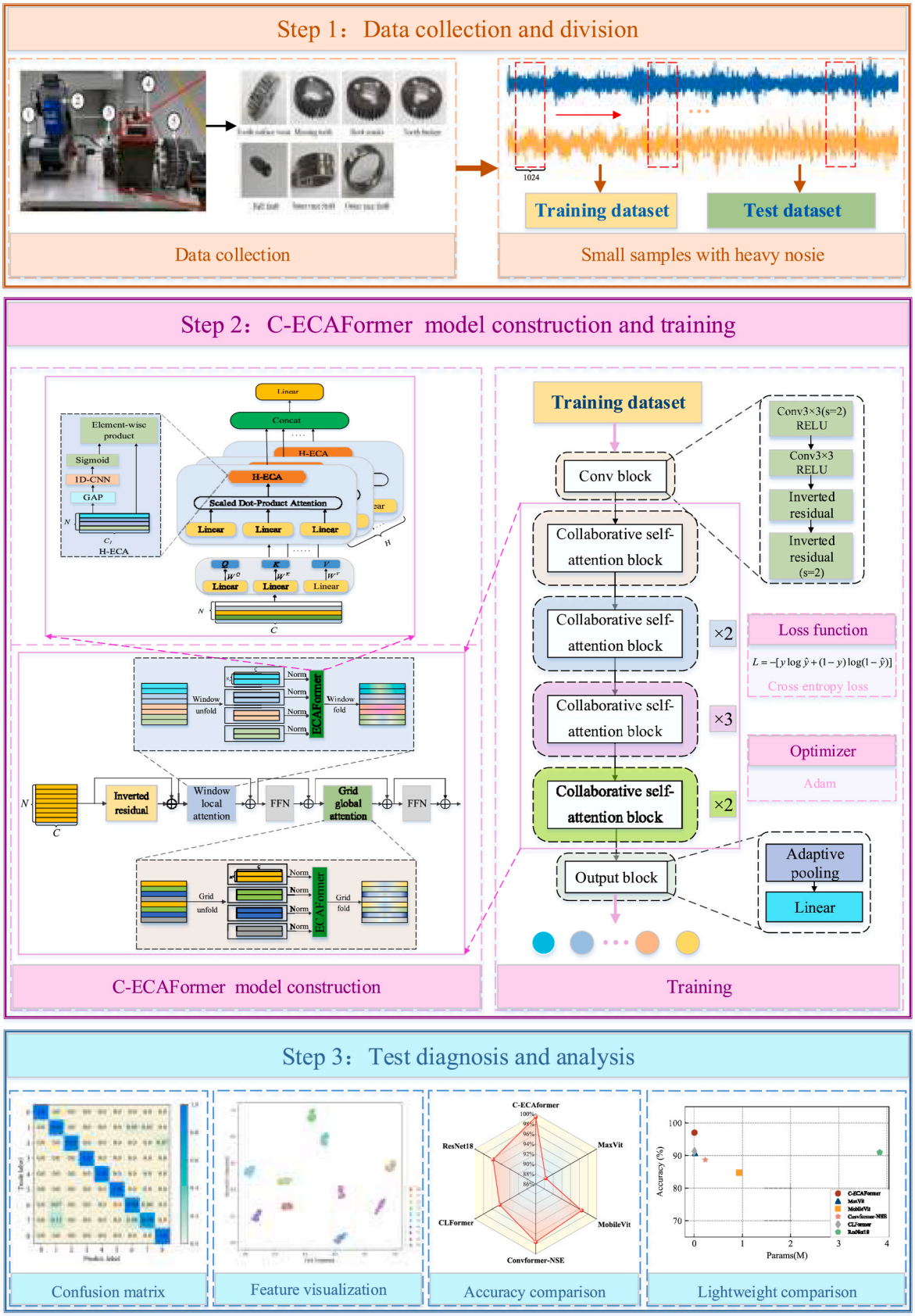


**Fig. 7.** C-ECAFormer based fault diagnosis method.

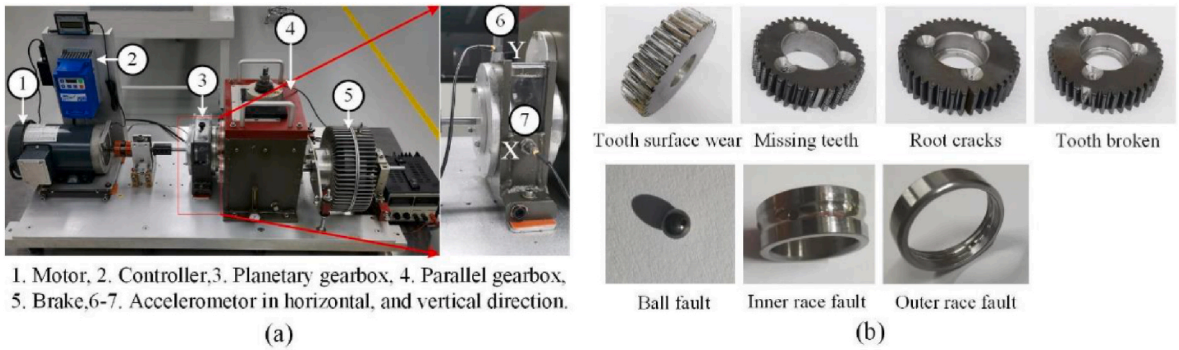


**Fig. 8.** The test bench of the XJTU Gearbox: (a) The test bench; (b) Fault modes of gears and bearings.

**Table 2**
Nine health conditions of XJTU Gearbox.

| Health conditions of planetary gearbox | Labels of conditions |
|---|---|
| Ball fault | 0 |
| Inner race fault | 1 |
| Mix fault (inner + out + ball) | 2 |
| Out race fault | 3 |
| Tooth broken | 4 |
| Missing tooth | 5 |
| Normal | 6 |
| Root cracks | 7 |
| Tooth surface wear | 8 |

During the training process, we record the accuracy and loss rate of the training and test sets for five trials. Fig. 9 shows the curves of the average accuracy and loss rate of the five trials with the number of iterations. From Fig. 9, it can be seen that C-ECAFormer, MaxVit, MobileVit and ResNet18 have a fast convergence rate and basically converge after 20 iterations, while Convformer-NSE and CLFormer converge after 70 iterations. C-ECAFormer tends to smooth out after convergence and does not fluctuate much, while the other three models with faster convergence still have significant fluctuations after convergence. At the same time, the accuracy of the converged C-ECAFormer is significantly higher than that of the other five models.

To further evaluate the robustness of the proposed model, we add noise to the input signals with SNR of -2dB, -6 dB, and −10dB, together with the initial experimental results with SNR of 2 dB. Table 3 presents the accuracy and complexity of each method under different SNR conditions, while Fig. 10 shows the bar chart of accuracy for each method under different SNR conditions. Through the analysis of the results, we

**Table 3**
Diagnostic accuracy and model complexity of each method under different levels of SNR.

| Methods | Accuracy (%) | | | | Complexity (M) | |
|---|---|---|---|---|---|---|
| | Different levels of SNR | | | | | |
| | 2 | −2 | −6 | −10 | Params | FLOPs |
| **C-ECAFormer** | **100** | **99.04** | **94.67** | **94.40** | **0.006** | **1.678** |
| CLFormer | 93.89 | 91.56 | 91.89 | 87.84 | 0.005 | 0.144 |
| Convformer-NSE | 97.20 | 89.22 | 87.53 | 80.82 | 0.230 | 6.255 |
| MaxVit | 98.82 | 98.37 | 88.44 | 77.44 | 0.017 | 6.510 |
| MobileVit | 97.78 | 95.33 | 78.67 | 67.22 | 0.933 | 54.554 |
| ResNet18 | 99.56 | 98.24 | 89.00 | 77.16 | 3.849 | 175.920 |

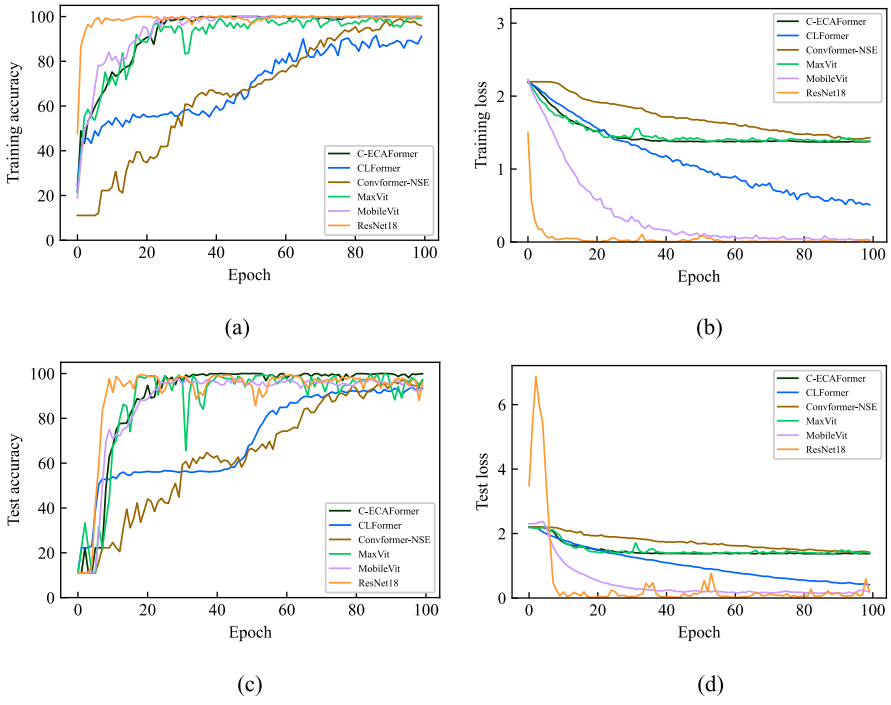


**Fig. 9.** The average accuracy and loss of each method: (a) Training accuracy; (b) Training loss; (c) Test accuracy; (d) Test loss.

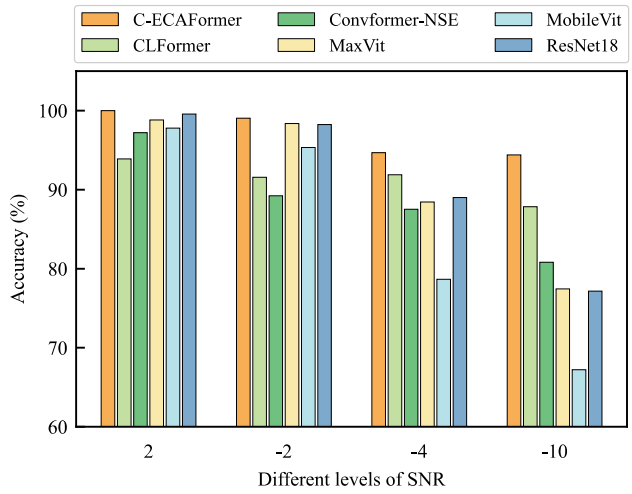


**Fig. 10.** Diagnostic accuracy of each method under different levels of SNR.

observe that as the noise increased, the computational accuracy of each method correspondingly decreases. Under SNR conditions of 2 dB and -2dB, the proposed model achieves accuracies of 100% and 99.04%, slightly higher than MaxVit's 98.82% and 98.37%, as well as ResNet18's 99.56% and 98.24%. However, under an SNR of -6dB, all the comparative methods exhibit significant accuracy declines, with accuracies lower than 90%, while the proposed method achieves an accuracy of 94.67%. When the noise further increased to a SNR of −10dB, the performance gap becomes more apparent. The accuracies of the comparative methods are all lower than 90%, while the proposed method still maintains high diagnostic performance with an accuracy of 94.40%. This is because C-ECAFormer enhances the relationship perception between different channels in the input feature map, enabling it to capture key information from important channels even under strong noise interference.

Fig. 11 shows the relationship between the average accuracy at four different SNR ratios and model complexity of the various methods for four different SNR. From the graph, we can observe that the proposed model achieves good diagnostic performance while simultaneously considering robustness and lightweight characteristics. Across the four SNR conditions, the average accuracy reaches 97.02%, with a model complexity of only 0.006M Params and 1.678M Floops. This effectively verifies that the designed Collaborative self-attention block can reduce the number of model parameters, improve computational efficiency, and reduce computational costs. Although CLFormer has a lower model complexity with 0.005M Params and 0.144M Floops, its average accuracy is 90.77%, which is 6.25% lower than the proposed model. The remaining four comparative models are inferior to the proposed model in terms of both accuracy and model parameters.

To further explore the feature extraction capability of the various methods under strong noise interference, a confusion matrix is used to analyse the classification details of the various methods under strong noise conditions with a SNR of −10dB. This is shown in Fig. 12. The coordinates 0–8 on the axes correspond to the nine health conditions on Table 2, respectively. As can be seen from the figure: the proposed method can detect most of the present health conditions, reaching accuracy over 93% in all except health conditions 6 and 7, where health conditions 0, 3, 4, 5 and 8 are identified with an accuracy of 100%. CLFormer achieves an accuracy of over 90% for health conditions 0, 1, 3, 4 and 5, but a lower accuracy of 48% for health condition 7, which is suboptimal. The remaining methods are significantly lower than the proposed method in terms of classification results. In addition, all methods except Convformer-NSE achieve lower results in identifying health condition 7 (Root cracks) than in identifying other health conditions to a greater or lesser extent, because the cracks in the root crack may be small or hidden inside the tooth root and the fault information is easily masked by strong noise.

### 4.2. Case 2: fault diagnosis of bearing towards heavy noise and small samples

#### 4.2.1. Data preparation of case 2

The CWRU dataset is provided by the Case Western Reserve University Bearings Data Center (Smith and Randall, 2015). The experimental platform is shown in Fig. 13. The experimental platform consists of a 2hp motor, a torque transducer, encoder, a dynamometer and control electronics. The experiments are carried out under four different motor loads and vibration signals are collected at 12 kHz or 48 kHz from normal bearings and damaged bearings with single point defects. Under each operating condition, single point faults are found on the rolling elements, inner and outer rings with fault diameters of 0.007, 0.014 and 0.021 inches respectively.

This case uses data collected at the drive end of the motor housing with a sampling frequency of 12 kHz, a motor load of 0 hp and a motor speed of 1797 rpm. Based on different fault sizes, the data is divided into 10 classes, including one class for a healthy bearing and three classes for each of the fault modes: inner race fault, rolling element fault, and outer race fault, with three different fault diameters. And for simple representation each working condition is given a different label, as shown in Table 4. A sliding window is used to divide the vibration signal, with each sample containing 1024 sampling points and a 30% overlap of two consecutive sampling points. For each health condition 50 training samples and 100 test samples are selected. To verify the robustness of the proposed method, Gaussian white noise with a SNR of -5dB is added to the input signal.

#### 4.2.2. Result analysis of case 2

In this experiment, the same five models as in Case 1 are selected as comparative experiments, and the experimental setup is the same as in

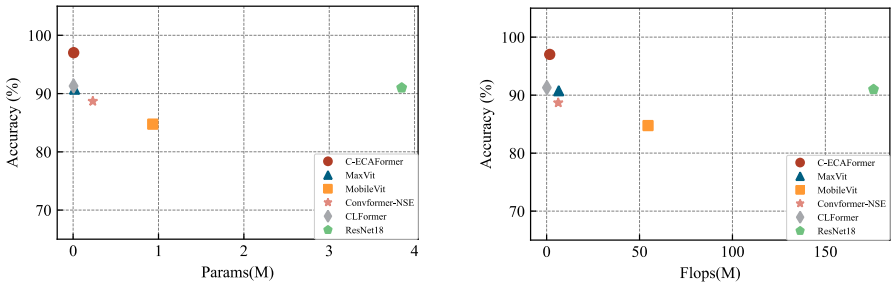


**Fig. 11.** Comparison of average accuracy and model complexity: (a) Performance of Params vs accuracy; (b) Performance of FLOPs vs accuracy.

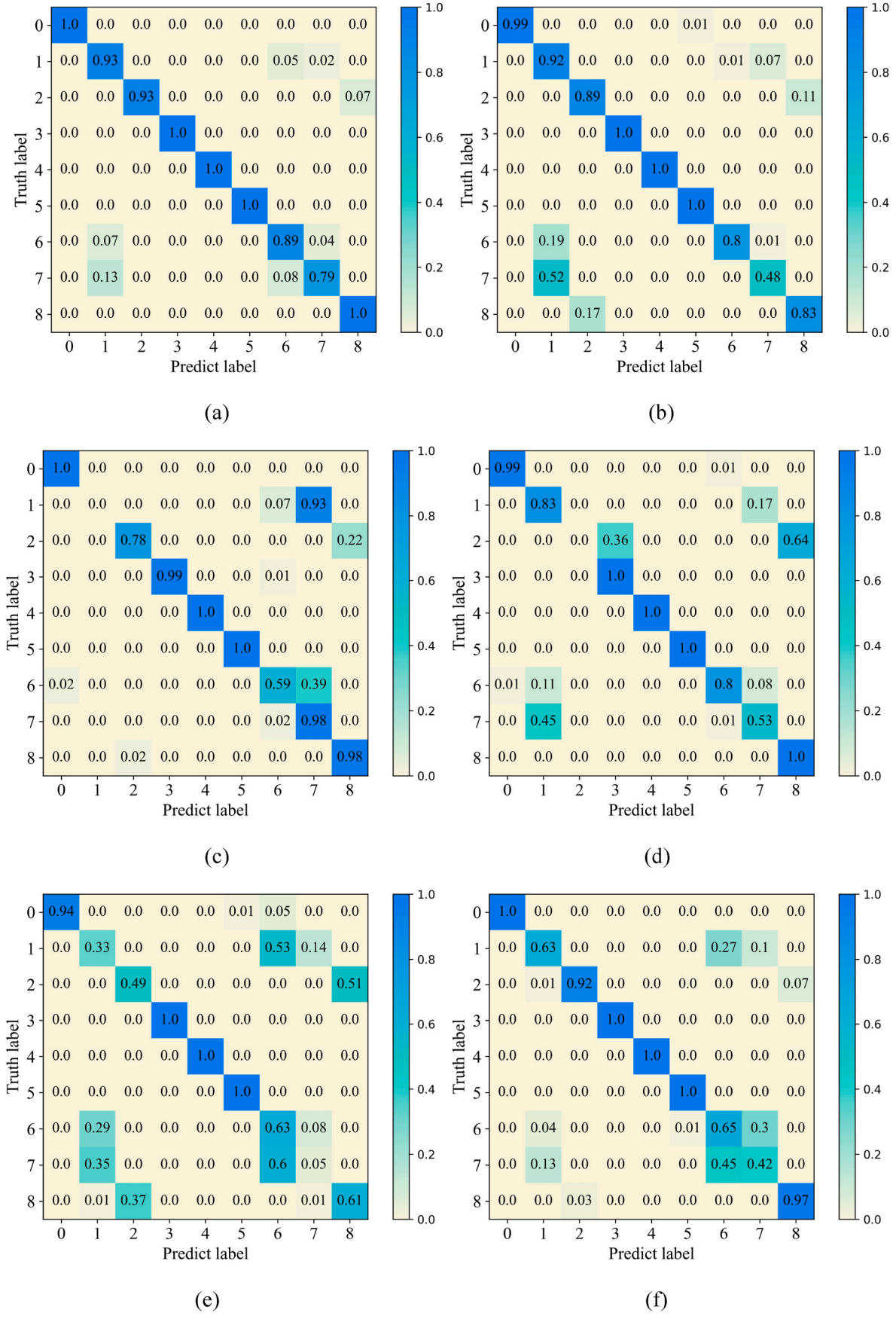


**Fig. 12.** Confusion matrix of the each method: (a) C-ECAFormer; (b) CLFormer; (c) Convformer-NSE; (d) MaxVit; (e) MobileVit; (f) ResNet18.

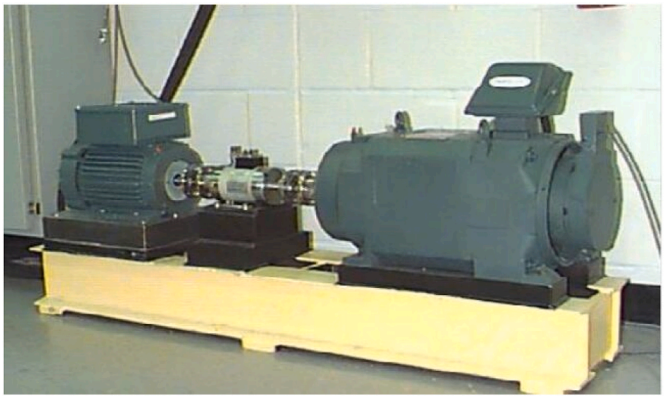

**Fig. 13.** The test bench of the CWRU bearing.

**Table 4**
Ten health conditions of CWRU bearing.

| Health conditions of bearing | Labels of conditions |
|---|---|
| Inner ring with 0.007 inch fault | 0 |
| Inner ring with 0.014 inch fault | 1 |
| Inner ring with 0.021 inch fault | 2 |
| Health State with 0.007 inch fault | 3 |
| Outer ring with 0.007 inch fault | 4 |
| Outer ring with 0.014 inch fault | 5 |
| Outer ring with 0.021 inch fault | 6 |
| Rolling Element with 0.007 inch fault | 7 |
| Rolling Element with 0.014 inch fault | 8 |
| Rolling Element with 0.021 inch fault | 9 |

Case 1, we repeated each set of experiments five times and then averaged the results of the five experiments.

In addition to accuracy, precision and recall are used as evaluation metrics for the models. Precision represents the proportion of predicted positive samples that are actually positive, measuring the false positive rate of the model. Recall represents the proportion of all positive samples that are correctly identified as positive, measuring the false negative rate of the model. The relevant formula is as follows:

$$Precision = \frac{TP}{TP + FN} \tag{13}$$

$$Recall = \frac{TP}{TP + FP} \tag{14}$$

where *TP* represents the samples that are actually positive and predicted as positive, *FN* represents the samples that are actually positive but predicted as negative, and *FP* represents the samples that are actually negative but predicted as positive.

Fig. 14 presents the average accuracy, precision, and recall of each method over five experiments. Analyzing the graph, it can be observed that the proposed method achieves an accuracy of 99.40%, precision of 99.41%, and recall of 99.40%, all of which are higher than the other comparative methods. Compared to the second-best performing method, Convformer-NSE, the proposed method outperforms it by 1.82% in accuracy, 1.68 in precision, and 1.82% in recall.

To explore the performance of each method in feature extraction, t-distributed stochastic embedding (t-SNE) is used for dimensionality reduction and feature visualization, as shown in Fig. 15 (Van der Maaten and Hinton, 2008). The visualization results demonstrate that the proposed method effectively forms clear classification boundaries, with dense clustering of different sample types and strong discriminative ability. In contrast, the other comparative methods exhibit overlapping of samples from different health conditions during the classification process. CLFormer, which performs second-best in terms of accuracy, shows slight overlap between health condition 0 (Inner ring with 0.007 inch fault) and 2 (Inner ring with 0.021 inch fault), 5 (Outer ring with 0.014 inch fault), 7 (Rolling Element with 0.007 inch fault), and 6 (Outer ring with 0.021 inch fault), 8 (Rolling Element with 0.014 inch fault).

To further explore the performance of the proposed method under small sample size, we conduct additional experiments with training set sizes of 30 and 10, in addition to the original experiment with a training set size of 50. These results in three sets of comparative experiments, as detailed in Table 5.

A box plot of the experimental results is shown in Fig. 16. Analysis of the Dataset 1 results shows that when the number of training samples is sufficient, the accuracy, precision and recall of all the models, except MaxVit, can reach over 94%. As the number of training samples decreases, the diagnostic performance of each method is affected to varying degrees. When the sample size is reduced to 30, the proposed method is less affected, with accuracy, precision and recall decreasing by 0.58%, 0.55% and 0.58% respectively. MobileVit is more affected by the sample size, with accuracy, precision and recall decreasing by 13.46%, 13.85% and 13.46%, respectively. The accuracy, precision and recall of the remaining methods decrease between 3% and 5%. When the number of training samples is further reduced to 10, significant differences in results between the methods emerge, Convformer-NSE and CLFormer are most affected, with accuracy dropping to 36% and 42%, respectively. MaxVit, MobileVit and ResNet18 are the next most affected, with accuracy rates of 74.42%, 67.6% and 68.5%. The proposed method is the least affected and still achieves 87.28% accuracy for a small sample. This is because the proposed method not only reduces the number of parameters and complexity of the model through the CSA block, but also pays more attention to the correlation between different channels, thus achieving better results with a signal-to-noise ratio of -5dB for small samples.

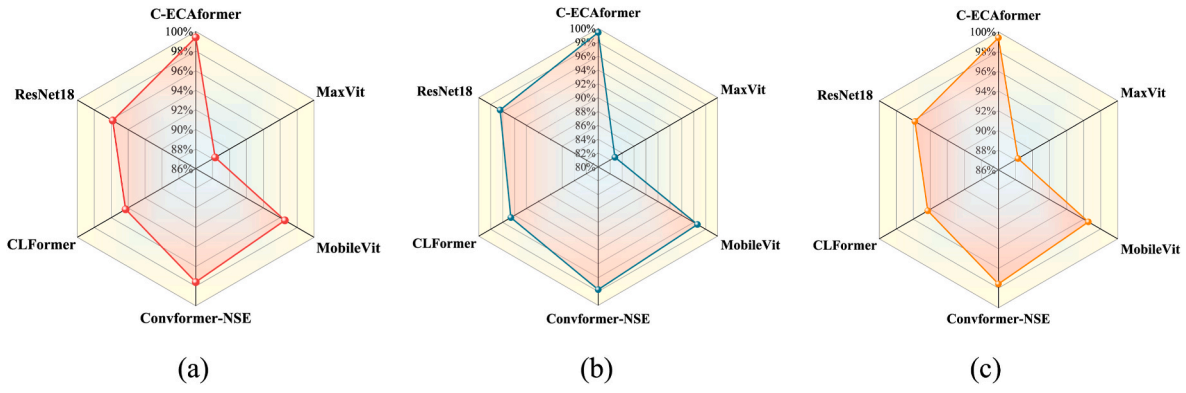


**Fig. 14.** Test results of each method: (a) Accuracy; (b) Precision; (c) Recall.

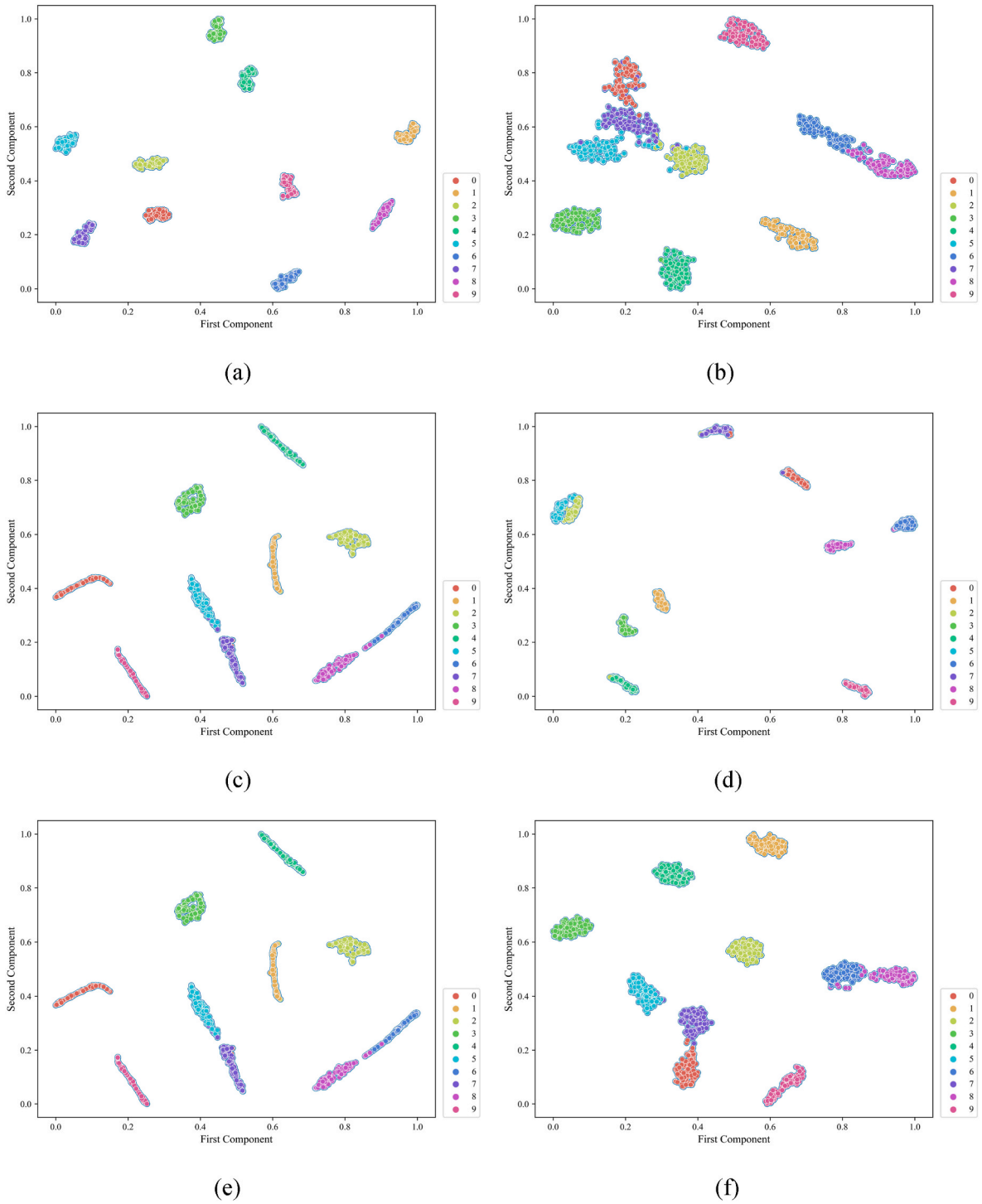


**Fig. 15.** Visualization of the extracted features: (a) C-EcaFormer; (b) CLFormer; (c) Convformer-NSE; (d) MaxVit; (e) MobileVit; (f) ResNet18.

**Table 5**
Detailed information on different experiments.

| Different datasets | Training sizes (Small) | Test sizes | SNRs of the added noise |
|---|---|---|---|
| Dataset 1 | 50 × 10 | 100 × 10 | −5 dB |
| Dataset 2 | 30 × 10 | 100 × 10 | −5 dB |
| Dataset 3 | 10 × 10 | 100 × 10 | −5 dB |

## 5. Conclusion

To improve the performance of intelligent fault diagnosis in scenarios with limited samples and heavy noise interference, this paper proposes a new lightweight model named as C-ECAFormer. Considering the computational complexity introduced by typical self-attention, a CSA block is designed. The block sequentially calculates window local attention and grid global attention, effectively reducing the computational cost while maintaining efficiency. During the attention calculation process in each block, we construct ECAFormer to capture fine-grained features of key channels and capture the interdependencies

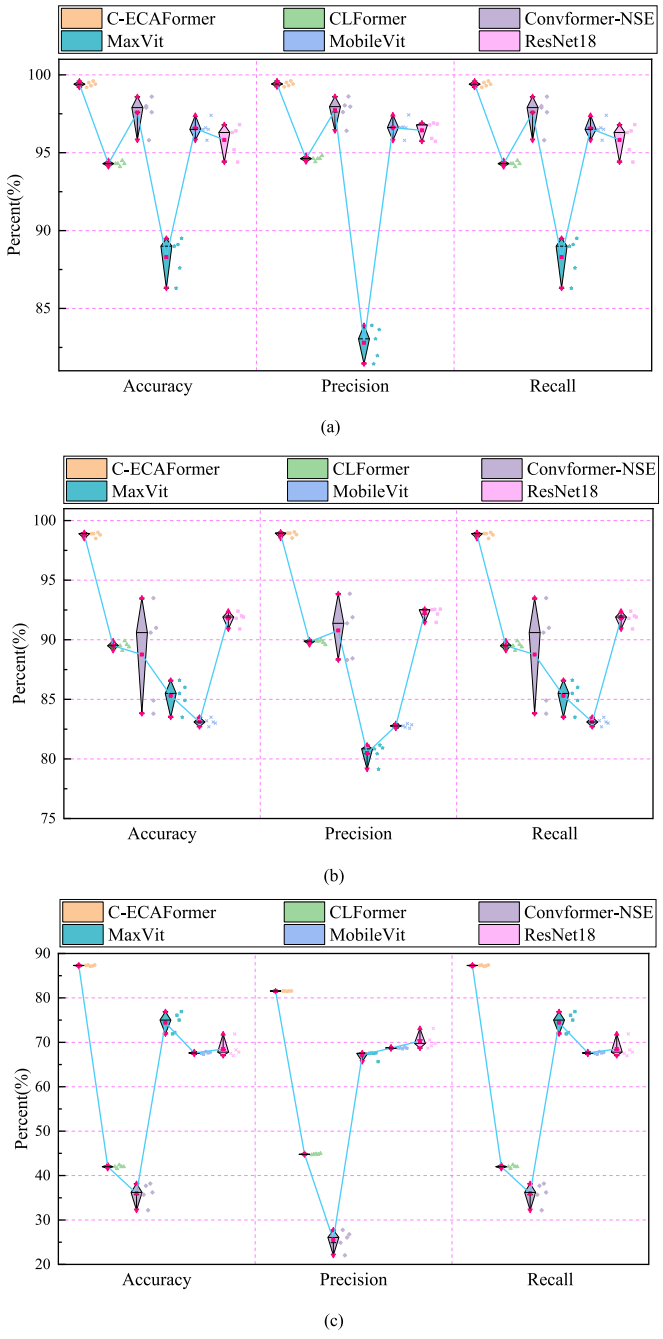


**Fig. 16.** Test results of each method: (a) Dataset 1; (b) Dataset 2; (c) Dataset 3.

between different channels. To leverage the strong capability of CNN in extracting low-level features, we introduce inverted residual block before the attention calculation to extract local information and modify the channel dimensions, thereby obtaining richer feature representations.

The experimental results from two distinct scenarios demonstrate the superiority of our proposed model compared to the existing mainstream Transformer and CNN models. The developed method exhibits robustness and lightweight characteristics while achieving higher accuracy in recognizing faults in situations involving limited samples and heavy noise interference. In future work, we will explore the applicability of the proposed method in various scenarios, aiming to enhance its effectiveness in dealing with more complex noise interference and even fewer samples.

## Declaration of competing interest

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.

## Data availability

Data will be made available on request.

## Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. 52275104) and the Natural Science Fund for Excellent Young Scholars of Hunan Province (No. 2021JJ20017).

## References

Chen, X., Shao, H., Xiao, Y., et al., 2023. Collaborative fault diagnosis of rotating machinery via dual adversarial guided unsupervised multi-domain adaptation network. Mech. Syst. Signal Process. 198, 110427.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al., 2020. An image is worth 16x16 words: transformers for image recognition at scale, pp. 1–22 arXiv preprint arXiv: 2010.11929.
Fang, H., Deng, J., Bai, Y., et al., 2021. CLFormer: a lightweight transformer based on convolutional embedding and linear self-attention with strong robustness for bearing fault diagnosis under limited sample conditions. IEEE Trans. Instrum. Meas. 71, 1–8.
Han, T., Liu, C., Yang, W., et al., 2019. A novel adversarial learning framework in deep convolutional neural network for intelligent diagnosis of mechanical faults. Knowl. Base Syst. 165, 474–487.
Han, S., Shao, H., Cheng, J., et al., 2023. Convformer-NSE: a novel end-to-end gearbox fault diagnosis framework under heavy noise using joint global and local information. IEEE-ASME Trans. Mechatron. 28 (1), 340–349.
He, K., Zhang, X., Ren, S., et al., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.
Hoang, D.T., Kang, H.J., 2019. Rolling element bearing fault diagnosis using convolutional neural network and vibration image. Cognit. Syst. Res. 53, 42–50.
Hou, Y., Wang, J., Chen, Z., et al., 2023. Diagnosisformer: an efficient rolling bearing fault diagnosis method based on improved Transformer. Eng. Appl. Artif. Intell. 124, 106507.
Jing, L., Zhao, M., Li, P., et al., 2017. A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox. Measurement 111, 1–10.
Lecun, Y., Bottou, L., Bengio, Y., et al., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.
Li, Y., Zhou, Z., Sun, C., et al., 2022a. Variational attention-based interpretable transformer network for rotary machine fault diagnosis. IEEE Transact. Neural Networks Learn. Syst. https://doi.org/10.1109/TNNLS.2022.3202234.
Li, T., Zhou, Z., Li, S., et al., 2022b. The emerging graph neural networks for intelligent fault diagnostics and prognostics: a guideline and a benchmark study. Mech. Syst. Signal Process. 168, 108653.
Mehta, S., Rastegari, M., 2021. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178.
Meng, H., Geng, M., Han, T., 2023. Long short-term memory network with Bayesian optimization for health prognostics of lithium-ion batteries based on partial incremental capacity analysis. Reliab. Eng. Syst. Saf. 236, 109288.
Sandler, M., Howard, A., Zhu, M., et al., 2018. Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520.
Shao, H., Zhang, X., Cheng, J., et al., 2020. Intelligent fault diagnosis of bearing using enhanced deep transfer auto-encoder. J. Mech. Eng. 56 (9), 84–91 (In Chinese).
Shao, H., Li, W., Cai, B., et al., 2022. Dual-threshold attention-guided Gan and limited infrared thermal images for rotating machinery fault diagnosis under speed fluctuation. IEEE Trans. Ind. Inf. https://doi.org/10.1109/TII.2022.3232766.
Smith, W.A., Randall, R.B., 2015. Rolling element bearing diagnostics using the Case Western Reserve University data: a benchmark study. Mech. Syst. Signal Process. 64, 100–131.
Sun, J., Gu, X., He, J., et al., 2022. A robust approach of multi-sensor fusion for fault diagnosis using convolution neural network. J. Dyn. Monit. Diagn. 103–110.
Sun, Z., Wang, Y., Gao, J., 2023. Intelligent fault diagnosis of rotating machinery under varying working conditions with global-local neighborhood and sparse graphs embedding deep regularized autoencoder. Eng. Appl. Artif. Intell. 124, 106590.
Tang, J., Zheng, G., Wei, C., et al., 2022. Signal-transformer: a robust and interpretable method for rotating machinery intelligent fault diagnosis under variable operating conditions. IEEE Trans. Instrum. Meas. 71, 1–11.
Tian, A., Zhang, Y., Ma, C., et al., 2023. Noise-robust machinery fault diagnosis based on self-attention mechanism in wavelet domain. Measurement 207, 112327.
Tu, Z., Talebi, H., Zhang, H., et al., 2022. Maxvit: multi-axis vision transformer. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October

23–27, 2022, Proceedings, Part XXIV. Springer Nature Switzerland, Cham, pp. 459–479.
Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11).
Vaswani, A., Shazeer, N., Parmar, N., et al., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
Wang, Q., Wu, B., Zhu, P., et al., 2020. ECA-Net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534–11542.
Wang, H., Xu, J., Yan, R., 2023. Intelligent Fault diagnosis for planetary gearbox using transferable deep Q network under variable conditions with small training data. J. Dyn. Monit. Diagn. 2 (1), 30–41.
Wu, Y., Zhao, R., Jin, W., et al., 2021. Intelligent fault diagnosis of rolling bearings using a semi-supervised convolutional neural network. Appl. Intell. 51, 2144–2160.
Xi, C., Yang, J., Liang, X., et al., 2023. An improved gated convolutional neural network for rolling bearing fault diagnosis with imbalanced data. Int. J. Hydromechatron. 6 (2), 108–132.
Xiao, Y., Shao, H., Feng, M., et al., 2023. Towards trustworthy rotating machinery fault diagnosis via attention uncertainty in transformer. J. Manuf. Syst. 70, 186–201.
Yao, J., Han, T., 2023. Data-driven lithium-ion batteries capacity estimation based on deep transfer learning using partial segment of charging/discharging data. Energy 271, 127033.
Zhang, W., Li, C., Peng, G., et al., 2018. A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. Mech. Syst. Signal Process. 100, 439–453.
Zhang, Y., Ji, J.C., Ren, Z., et al., 2023a. Digital twin-driven partial domain adaptation network for intelligent fault diagnosis of rolling bearing. Reliab. Eng. Syst. Saf. 234, 109186.
Zhang, Y., Ren, Z., Feng, K., et al., 2023b. Universal source-free domain adaptation method for cross-domain fault diagnosis of machines. Mech. Syst. Signal Process. 191, 110159.
Zhao, R., Yan, R., Wang, J., et al., 2017. Learning to monitor machine health with convolutional bi-directional LSTM networks. Sensors 17 (2), 273.
Zhen, D., Li, D., Feng, G., et al., 2022. Rolling bearing fault diagnosis based on VMD reconstruction and DCS demodulation. Int. J. Hydromechatron. 5 (3), 205–225.
Zhou, K., Tong, Y., Li, X., et al., 2023. Exploring global attention mechanism on fault detection and diagnosis for complex engineering processes. Process Saf. Environ. Protect. 170, 660–669.