Measurement Science and Technology

100 year anniversary
Measurement Science and Technology

PURPOSE-LED PUBLISHING™

PAPER

# Bearing fault diagnosis using joint features extraction with multi-scale residual convolutional neural network and transformer

To cite this article: Shuzhen Han *et al* 2025 *Meas. Sci. Technol.* **36** 056108

View the article online for updates and enhancements.

IOP Publishing Measurement Science and Technology

Meas. Sci. Technol. **36** (2025) 056108 (20pp) https://doi.org/10.1088/1361-6501/adc761

# Bearing fault diagnosis using joint features extraction with multi-scale residual convolutional neural network and transformer

**Shuzhen Han**[1,5], **Jianfei Li**[2], **Ke Pang**[2], **Dong Zhen**[3], **Guojin Feng**[3], **Fujun Tian**[4] **and Pingjuan Niu**[1,2,*]

[1] School of Mechanical Engineering, Tiangong University, Tianjin 300387, People's Republic of China
[2] School of Electronics and Information Engineering, Tiangong University, Tianjin 300387, People's Republic of China
[3] School of Mechanical Engineering, Hebei University of Technology, Tianjin 300401, People's Republic of China
[4] School of Artificial Intelligence, Tiangong University, Tianjin 300387, People's Republic of China
[5] Office of Cyberspace Affairs of Tiangong University, Tiangong University, Tianjin 300387, People's Republic of China

E-mail: niupingjuan@tiangong.edu.cn, hanshuzhen@tiangong.edu.cn, lijianfei@tiangong.edu.cn, pangke@tiangong.edu.cn, D.Zhen@hebut.edu.cn, G.Feng@hebut.edu.cn and tianfujun@tiangong.edu.cn

Received 15 January 2025, revised 25 March 2025
Accepted for publication 31 March 2025
Published 23 April 2025

**Abstract**

In recent years, deep learning has shown significant potential in bearing fault diagnosis. However, challenges remain, including suboptimal signal quality under intricate conditions and the impact of network architecture on model performance. This study proposes a joint feature extraction method that combines a multi-scale residual convolutional neural network with a position-encoded transformer and integrating a transfer learning (TL) strategy to address the aforementioned issues. First, multi-scale convolutional layers are utilized to capture the details and local features from raw signals. To complement this, a transformer with embedded positional encoding learns global dependencies while retaining position information, effectively compensating the position shift issue. To further enhance the model's generalization capability, a data augmentation strategy is designed and implemented, diversifying the training data. Furthermore, a TL strategy predicated on model fine-tuning is applied to mitigate the reliance on a substantial quantity of labeled data. Experiments conducted on two datasets, including Case Western Reserve University (CWRU) dataset and Self-Collected two-stage gear drive test bench dataset, that featuring diverse working conditions and bearing types. The proposed model achieved 99.92% accuracy on the CWRU dataset and demonstrated strong robustness across different operating conditions. With a model size of approximately 1.45MB, it maintains high diagnostic performance while ensuring computational efficiency. This intelligent diagnosis technique can also be applied to other rotating machinery, such as wind power, locomotive and various other fields owing to its robust feature extraction abilities.

Keywords: bearing fault diagnosis, MSRCNN, transformer, transfer learning

[*] Author to whom any correspondence should be addressed.

## Abbreviations

| | |
|---|---|
| CNN | Convolutional neural network |
| CWRU | Case Western Reserve University |
| DANN | Domain-adversarial neural network |
| DCNN | Deep convolutional neural network |
| DL | Deep learning |
| ECTN | Efficient convolutional transformer |
| FCL | Fully-connected layer |
| FFN | Feed-forward network |
| HP | Horsepower |
| MHA | Multi-head attention |
| MLP | Multi-layer perceptron |
| MSA | Multi-head self-attention |
| MSRCNN | Multi-scale residual convolutional neural network |
| MSRCT | Multi-scale residual convolutional neural network and transformer |
| PE | Positional encoding |
| QKV | Query-key-value |
| ReLU | Rectified linear unit |
| SIT | Signal transformer neural network |
| SNR | Signal-to-noise ratio |
| TICNN | Time invariant convolutional neural network |
| TL | Transfer learning |
| TL-TAR | Transfer learning-transformer and ResNet |
| ViT | Vision transformer |
| WDCNN | Wide deep convolutional neural network |
| WDD-CNN | Wide and deep dilated convolutional neural network |

## 1. Introduction

Bearings, as critical components in rotating machinery, significantly impact overall equipment reliability [1, 2]. Their failures can result in substantial production losses and safety hazards, thus highlighting the necessity of reliable fault diagnosis [3]. Vibration-based condition monitoring is one of the most effective methods for early fault detection, and feature extraction plays a crucial role in analyzing vibration signals. A comprehensive review by Riaz *et al* [4]. systematically summarizes various vibration feature extraction techniques and their impact on fault diagnosis of rotating machinery, emphasizing the importance of time-domain, frequency-domain, and time–frequency domain analysis. The integration of these features with machine learning has substantially contributed to the advancement of intelligent fault diagnosis by enhancing pattern recognition capability and improving diagnostic accuracy under complex conditions. With the advent of Industry 4.0, traditional diagnostic methods have proven inadequate for meeting the demands of intelligent manufacturing [5]. Furthermore, the quality of bearing vibration signals is often compromised by various factors, including load fluctuations and complex environments, presenting significant challenges in fault diagnosis under complex working conditions [6].

In recent years, DL has driven extensive exploration in the field of end-to-end intelligent fault diagnosis. Owing to its remarkable capacity to discern features directly from raw data, DL has already emerged as a highly promising methodology for bearing fault diagnosis [7]. CNNs, a DL model specifically designed to construct network structures, have demonstrated outstanding performance and effectiveness in the field of fault diagnosis [8]. It provides an effective method of directly processing raw signals without requiring manual preprocessing [9]. A previous study [10] leveraged CNN to integrate the benefits of image analysis and visual perception into the process of diagnosing bearing faults. Although CNN has shown good performance in image recognition, its application in industrial settings remains limited. In recent years, in order to break through this dilemma, some scholars have attempted to improve CNN to make it more suitable for industrial applications and further exploring their potential in the industrial field. Liu *et al* [11] proposed a misaligned time-series CNN architecture that addresses the limitations of traditional CNN in extracting features from periodic mechanical signals. Wen *et al* [12] proposed a method to convert signals into two-dimensional images, which enables the extraction of the features inherent to the converted images, thereby negating the influence of handcrafted features. The results demonstrate that the proposed fault diagnosis method has significant improvements. Yao *et al* [13] developed an end-to-end CNN is capable of processing time and frequency domain signals as raw inputs for the model and fusing multi-channel information from different sound signals through its channels without an extra fusion algorithm. In another study [14], vibration spectrum imaging was used as the input of the CNN to classify bearing faults, demonstrating excellent performance in the classification task.

The aforementioned studies indicate that CNN-based methods enable end-to-end fault diagnosis by automatically extracting fault features and eliminating the need for complex signal processing and expert knowledge. These methods have already demonstrated significant effectiveness. In practical applications, the acquisition of vibration signals often faces complex operating conditions, requiring the extraction of fault-related features for effective fault diagnosis. Despite advancements in CNN-based methods that facilitate automated feature extraction and various improvements, CNNs remain inherently constrained by their reliance on local receptive fields. This limitation restricts their ability to model long-range dependencies in vibration signals, and variable conditions continue to challenge their diagnostic accuracy and efficiency. Ongoing research thus focuses on optimizing deep learning architectures to contend with real-world complexities more effectively.

Researchers have sought to enhance feature extraction capabilities by deepening network architectures. Pan *et al* [15] learned features from the fused data through DCNN and made predictions via a fully connected layer, achieving a high fault diagnosis accuracy with reduced model training time. Zhang *et al* [16] proposed WDCNN that contains five layers of convolution and pooling. It employs a wide kernel in the first convolution layer and smaller kernels in subsequent layers for fault diagnosis, achieving high accuracy. Based on WDCNN, Zhang *et al* [17] proposed TICNN, using dropout, small batch training and other tricks to improve the ability of anti-noise and adaptation. The constantly changing loads of electric motors and the presence of noise can degrade the performance of diagnostic methods. Recent research [18] proposed WDD-CNN, a model combining one-dimensional and two-dimensional DCNNs with broad initial layer kernels. WDD-CNN can operate in real time and conduct early detection of potential faults. Nevertheless, deepening network architectures, while enhancing feature extraction capabilities, can introduce issues such as vanishing gradients, exploding gradients, and overfitting. These complications often lead to performance deterioration and significant disparities in model outcomes [19].

Building on the previous research, researchers have introduced residual blocks into CNN architectures to specifically address problems like vanishing and exploding gradients, thereby facilitating the development of deeper and more robust networks. An experimental study in the early stages proposed a residual learning architecture, exceptionally tailored for handling mechanical vibration signals characterized by variable sequence lengths [20]. Yan *et al* [21] proposed a method that converts vibration signal samples into two-dimensional images and establishes a deep residual neural network for feature extraction, achieving better average accuracy. To improve the accuracy of fault diagnosis, Yu *et al* [22] constructed a residual neural network with six layers (ResNet06), incorporating two residual units to comprehensively capture the characteristics of mechanical vibration signals. In the case of unexpected catastrophic failures, Hu *et al* [23] proposed a prior-knowledge-based residual shrinkage prototype network to address the challenges of fault diagnosis with limited labeled samples.

Although the aforementioned architectural enhancements have demonstrated notable improvements in feature extraction and classification performance, they inevitably introduce increased structural complexity into CNN-based models. With the growing depth and intricacy of network architectures, CNNs tend to become more sensitive to spatial variations in the input domain. The classification accuracy of CNNs may fluctuate considerably in response to shifts in the positional distribution of input signals. In contrast, the transformer can span the regions of chaotic local features caused by the offset of signal points, understand and process the signals as a whole, and solve the problems brought to classification by the signal point offset to a certain extent. Consequently, the popularity of transformer modules has grown considerably in recent times, largely due to the highly effective structure they offer [24]. Transformer demonstrates outstanding performance in pattern recognition by its unique architecture, such as the MHA mechanism. It can effectively capture the relationships between distant elements in a sequence through the self-attention mechanism [25]. Ahmad *et al* [26] proposed a pioneering multiple-attention model, MANS-Net. By leveraging channel attention, spatial attention, and Transformer-based attention modules, this model improves the precision of nuclei segmentation. With the transformer demonstrating its unique advantages in numerous fields, its application in the field of fault diagnosis has also attracted much attention. Yang *et al* [27] proposed a bearing fault diagnosis method based on the SIT. By using the transformer to segment, encode, and extract features from one-dimensional vibration signals, combined with a MLP to achieve fault recognition. The results of the experimental study demonstrate that transformer is highly accurate and performs exceptionally well in diagnostic applications. Tang *et al* [28] introduced an integrated ViT framework that leverages wavelet transformation and a soft consensus approach, improving diagnostic accuracy and generalization capabilities. Liu *et al* [29] proposed the ECTN for intelligent bearing fault diagnosis, deriving diagnostic results from a classifier after applying time–frequency representations to raw signals. Jiao *et al* [30] developed a new end-to-end fault diagnosis framework based on a binary tree filter transformer to enhance the diagnostic performance of traditional DL-based bearing fault diagnosis models. To strengthen the translation invariance capability of CNNs, a transformer equipped with PE is introduced during feature extraction. The PE effectively compensates for the transformer's inherent insensitivity to spatial positional information. Thus, when extracting fault features, the optimal global features can be obtained while retaining the positional information of the features.

Although improving the network model can enhance the feature extraction capability, the training process requires a substantial number of labeled fault samples [31]. However, in practical applications, obtaining a sufficient number of fault samples requires significant of human and material resources [32]. Therefore, the decline in diagnostic performance under different working conditions brings new challenges. TL stands as an efficacious strategy, leveraging the knowledge derived from a source domain dataset to address distinct yet pertinent tasks within a target domain dataset. In recent years, TL research [33–35] has yielded remarkable results in fault diagnosis. For example, some scholars [36] also proposed a bearing fault diagnosis method TL-TAR based on the joint feature extraction of TAR combined with a TL strategy. TL-TAR aims to address challenges such as limited data availability and the poor performance of existing DL models. However, these current methods have certain limitations at the ideological level, which hinder their overall effectiveness and performance in dealing with such challenges. The DANN proposed by Ganin *et al* [37] represents a novel domain adaptation method for representation learning. DANN employs a neural network architecture, training with labeled data in the source domain and unlabeled data in the target domain, then adding a gradient reversal layer, opening up new directions for TL research in the field of fault diagnosis.

Based on this comprehensive review, it is evident that while existing models have demonstrated notable diagnostic successes under certain controlled conditions, they exhibit considerable performance declines when confronted with challenges such as complex signal environments, variable operational conditions, and limited sample sizes. The conventional deep learning model, CNN, effectively automates feature extraction from raw signals but necessitates structural optimization to process intricate signals adeptly. In an innovative approach, the SIT model employs a self-attention mechanism extensively to adeptly capture the complex interdependencies among signal sequences, addressing the limitations of traditional models in managing intricate dependencies within signal data. However, its efficacy markedly diminishes with changes in operational conditions. The DANN model enhances cross-domain adaptability by incorporating a domain classifier that extracts domain-invariant features, yet its diagnostic capabilities under intense noise conditions remain constrained. The TL-TAR model employs a fusion architecture for feature extraction and improves adaptability to transfer tasks but overlooks the influence of fault categories and restricts fault feature extraction to a singular scale. Consequently, diagnosing bearing faults in complex scenarios continues to pose a significant challenge.

In response to the aforementioned problems, a new model named MSRCT is proposed which combines MSRCNN and Transformer with PE based on TL. The method eliminates the need for signal preprocessing or manual extraction of characteristic parameters. It automatically learns relevant features by processing raw vibration signals through the MHA mechanism and other means. Furthermore, MSRCNN focuses on local feature extraction, while transformer specializes in global feature integration and correlation analysis. The two components operate synergistically, realizing comprehensive fault feature learning and diagnosis, spanning from local details to overall trends. Meanwhile, the introduction of the TL strategy enables rapid training under specific conditions and resolves the problem of differences in the data characteristic distributions. Specifically, the key contributions of this paper are summarized as follows:

(i) The innovative feature extraction mechanism, designed to address the periodicity of vibration signals, integrates multi-scale analysis with a hybrid framework of residual connections and a transformer. Multi-scale analysis captures detailed signal information across various scales. The residual connections combat the vanishing gradient problem, ensuring robust transmission of information through the network layers. Concurrently, the transformer's MHA mechanism globally discerns complex signal relationships, which enhances the effectiveness of feature extraction for more precise fault diagnosis.

(ii) TL with fine-tuning is leveraged to enable efficient adaptation of the proposed model to target tasks, particularly in scenarios with limited labeled samples. By inheriting knowledge from the source domain and fine-tuning to align with target data characteristics, the approach improves accuracy and efficiency, reduces the dependence on a large amount of labeled data, and enhances the adaptability of the model in new domains.

(iii) To validate the robustness and generalizability of the proposed method, experiments wereperfommed under diverse operating conditions. These experiments utilized datasetsencompassing a wide range of operational conditions and bearing varieties. The resultsdemonstrate that the proposed method achieves high classification accuracy and exhibitsexcellent robustness.

## 2. Theoretical background

### 2.1. Convolutional operation

The convolutional operation involves sliding a small convolutional kernel over the input data. At each position, it calculates the sum of the products of the elements at the corresponding positions of the convolutional kernel and the input data, and ultimately obtains an element in the output feature map. The process can be regarded as extracting local features from the input data. By choosing convolutional kernels of different sizes, local features of different scales can be captured. Small convolutional kernels focus on short-term local changes, whereas large convolutional kernels can capture longer-term trends and features.

Let the input sequence be $x = [x_1, x_2, \cdots, x_n]$, and the convolutional kernel be $k = [k_1, k_2, \cdots, k_m]$, where $n$ is the length of the input sequence, $m$ is the length of the convolutional kernel, and $m < n$. At a certain sliding position $i(1 \leqslant i \leqslant n - m + 1)$, the output $y_i$ of the convolutional operation is computed as follows:

$$y_i = \sum_{j=1}^{m} k_j x_{i+j-1} \tag{1}$$

The process can be regarded as the convolutional kernel 'scanning' over the input sequence to detect specific local structures or features.

### 2.2. Residual connection

The core idea behind the residual connection is the introduction of a shortcut connection within the network, enabling a certain layer of the network to directly access the input information of the previous layer. Specifically, for a network layer with a residual connection, its output consists of both the result of the normal transformation of the input and the original input itself. Let $x$ be the input to the network layer, and $H(x)$ be the output after the normal transformation, which involves operations such as convolution and activation. The output $y$ after the residual connection is then given by:

$$y = H(x) + x \tag{2}$$

In practice, for the convenience of training and optimization, the residual connection is usually expressed as learning the residual mapping $f(x) = H(x) - x$. In this case, the goal of the network becomes to learn the residual mapping $f(x)$, and the final output is given by $y = f(x) + x$.

The key advantage of incorporating residual connections within the network architecture is manifested as follows.

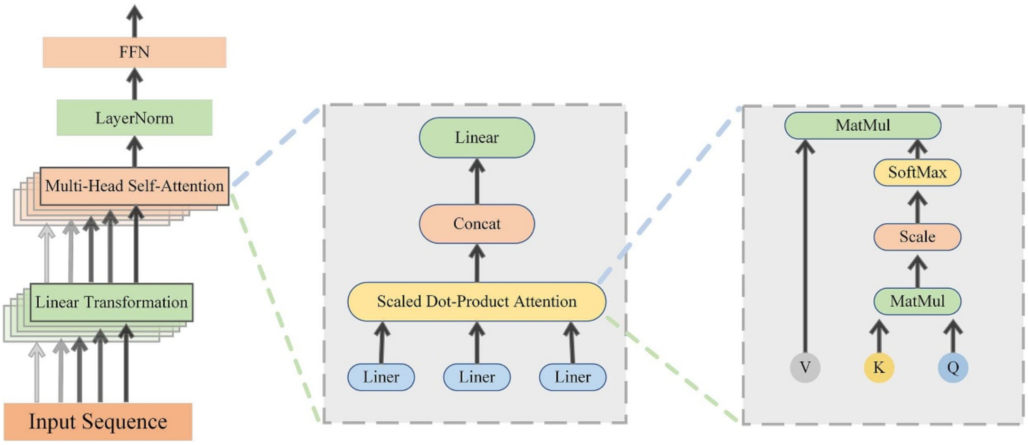


**Figure 1.** Transformer encoder.

As the network's depth escalates, in the event that the normal transformation $H(x)$ encounters issues like vanishing or exploding gradients, optimization becomes arduous.

The network is still capable of guaranteeing the efficient conveyance of information and maintaining the model's performance by learning a relatively smaller residual mapping $f(x)$.

### 2.3. Transformer encoder

The transformer encoder consists of a stack of multiple encoder layers with the same structure but different parameters. Through the successive stacking of layers, the model is capable of incrementally extracting increasingly sophisticated and abstract feature representations, thereby improving its performance in various natural language processing tasks. Each encoder layer contains two key components: the MSA mechanism and the FFN. To help mitigate the degradation phenomenon of deep networks, after the training of each component within the encoder layer has been successfully completed, residual connections and normalization operations will be performed. The transformer encoder can process each position of the input sequence in parallel, significantly enhancing the computational efficiency. Figure 1 displays the schematic diagram of the transformer encoder model.

*2.3.1. Principle of MSA mechanism.* The basic idea is to establish the interactions between different position elements in the input sequence, enabling effective capture of global information and feature extraction. For the input sequence $x$, first obtain the query vector $q$, key vector $k$, and value vector $v$ through linear mapping. We assuming the input sequence $x \in \mathbb{R}^{D_k}$, and $W_q$, $W_k$, $W_v$ are the parameter matrices of the linear mapping. The formulas are as follows:

$$Q = W_q^T X = [q_1, \cdots, q_N] \quad (3)$$

$$K = W_k^T X = [k_1, \cdots, k_N] \quad (4)$$

$$V = W_v^T X = [v_1, \cdots, v_N] \quad (5)$$

Next, the attention weights are computed by calculating the correlation scores between the query and key vectors using the scaled dot-product formula. Then, carry out the normalization process using the softmax function to obtain the attention distribution. Finally, perform a weighted sum of the value vectors according to the attention distribution to obtain the final attention output.

The MSA mechanism effectively captures global information in sequences through linear mapping and attention weight computation, making it suitable for processing long sequences. Specifically, the computation process of MSA mainly consists of two parts: the linear mapping part has a time complexity of $O(N \cdot D_k^2)$, where $N$ represents the sequence length and $D_k$ denotes the dimension of each attention head. This part is responsible for mapping the input sequence to the query, key, and value spaces. The attention part has a time complexity of $O(N^2 \cdot D_k)$ , primarily used to compute the attention weights between each pair of elements in the sequence. Therefore, the overall time complexity of MSA is $O(N \cdot D_k^2 + N^2 \cdot D_k)$. As the sequence length $N$ increases, the computational cost will rise. Although the computational cost of MSA is higher than traditional methods such as RNNs and CNNs, the advantage of MSA lies in its ability to better model long-range dependencies, thereby achieving a good balance between capturing global information and computational efficiency.

*2.3.2. Role of the FFN.* The FFN performs additional nonlinear transformation of features in each encoder layer. It is a straightforward two-layer network. For each position vector $z$ of the input, the FFN first applies a linear transformation with parameter $W_1$, then uses the ReLU activation function for non-linear activation, and then undergoes another linear transformation with parameter $W_2$. The FFN is mathematically defined as:

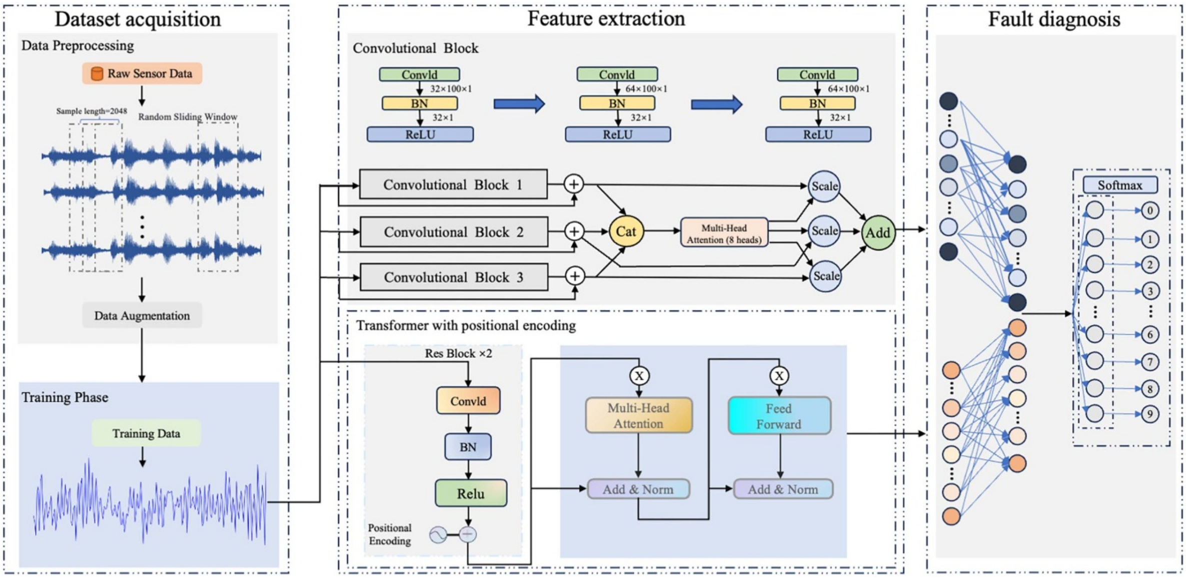


**Figure 2.** The framework of MSCRT.

$$\mathrm{FFN}\left(z\right) = W_2 \times \mathrm{ReLU}\left(W_1 z + b_1\right) + b_2 \tag{6}$$

where $b_1$ and $b_2$ are bias terms. The FFN increases the non-linear expression ability of the model and facilitates the learning of complex feature relationships.

## 3. Proposed fault diagnosis method

This section begins by detailing the framework of the proposed method, including the structure, component relationships, and the operation principle. Subsequently, the training strategies designed and applied for the bearing fault diagnosis method are thoroughly introduced.

### 3.1. Proposed framework

The proposed method consists of three key components: dataset acquisition, feature extraction, and fault diagnosis. The diagrammatic representation of the proposed methodology is presented in figure 2.

In the data processing procedure, a data augmentation operation is first applied to the original data. Specifically, a random sliding window sampling method is used to obtain samples, which are then designated as training data. Before feeding the data into the network, a specific standardization procedure is performed on the samples. The aim is to make the data conform to the characteristics of a standard Gaussian distribution, with a mean of 0 and a variance of 1. Subsequently, the acquired training data is fed into the MSRCNN module and the transformer module with PE for feature extraction. These two modules operate in parallel, concurrently extracting local and global features across multiple dimensions. They supplement one another and function in a synergistic manner, thereby significantly augmenting the feature extraction capabilities of the overall system. The extracted features are then concatenated and processed by the classification module. Finally, the classification results are attained through the Softmax function. The detailed information of each part will be illustrated in the following subsections.

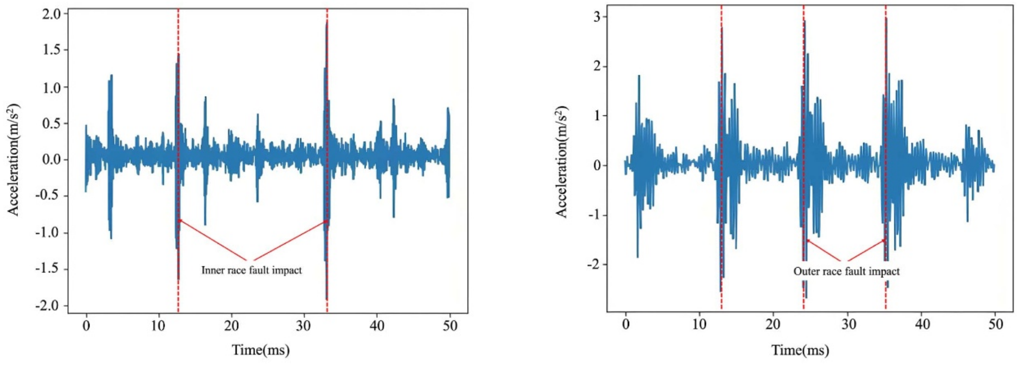
### 3.2. Multi-scale residual convolutional module

In complex mechanical systems, different categories of faults often generate impact signals with distinct characteristics. These impact signals contain valuable fault information. For example, in rotating machinery, a bearing fault may induce periodic impacts, as illustrated in figure 3, while a bearing fault may generate an impact associated to the bearing meshing frequency.

These impacts display distinctive patterns in both the time domain and the frequency domain. As the severity of the fault increases, the intensity and frequency characteristics of the impact signals will also change correspondingly.

To accurately capture the characteristics embedded in the impact signals generated by different faults and thus achieve accurate and reliable fault diagnosis, we introduce the method of Multi-scale convolution operation. Multi-scale convolution analyzes the input data from different resolution levels, akin to observing an object through magnifying glasses of varying magnifications. Multi-scale strategy can capture the details and structural information in the data more comprehensively.

We carefully designed a network structure composed of three convolutional layers, with the aim of performing in-depth feature extraction on the input data from multiple different scales. The process of the three convolutional layers is to perform three convolutional operations on the input data in turn. The first layer uses a smaller convolution kernel to capture local features, the second layer uses a slightly larger convolution kernel to expand the feature capture range, and the third layer uses the largest convolution kernel to obtain macroscopic features, and finally obtains the output feature map



(a) Inner race fault impact

(b) Outer race fault impact

**Figure 3.** The periodic impact caused by bearing fault.

integrating features of different scales. We can represent the three-layer convolution operation as a composite function. The specific formula is as follows:

$$F(X) = f_3\left(f_2\left(f_1\left(X\right)\right)\right) \tag{7}$$

where

$$f_1\left(X\right) = Y_1 = \mathrm{Conv}\left(X, k_1, c_1, s_1\right) \tag{8}$$

$$f_2\left(Y_1\right) = Y_2 = \mathrm{Conv}\left(Y_1, k_2, c_2, s_2\right) \tag{9}$$

$$f_3\left(Y_2\right) = \mathrm{Conv}\left(Y_2, k_3, c_3, s_3\right) \tag{10}$$

where $k$ represents the size of the convolution kernel, $c$ represents the number of convolution channels, and $s$ is the convolution stride.

To solve the problem of vanishing gradients in deep networks, we introduce a new strategy, that is, adding residual connections after the convolution output. During the backpropagation process, gradients can be directly transmitted to the shallow layers of the network through the residual connections, ensuring that the shallow layers can be effectively trained. Meanwhile, the residual connections are beneficial for the network to better learn useful features from the input data, because they can more easily capture the change information of the input data after the convolution operation. The specific operations are as follows:

$$Z = \mathrm{Output} + x \tag{11}$$

where $x$ is the input, and Output is the output after the convolution operation. The most crucial aspect of this connection method lies in prompting the network to learn from residual information, that is, the difference between Output and $x$.

To effectively fuse the multi-scale features obtained from the three-layer convolution and thus extract more robust, typical, and effective features, we employ a comprehensive feature fusion strategy. Specifically, we first apply a MHA mechanism to each of the feature maps generated by the three-layer convolution. The MHA mechanism can be formulated as follows. Given an input feature map $X$, we first project $X$ into query $Q$, key $K$ and value $V$ matrices via linear transformations. Then, we split the $Q$, $K$ and $V$ matrices into $h$ heads. For the $i$th head, the attention score is calculated as:

$$Attn_i = \mathrm{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \tag{12}$$

where $d_k$ is the dimension of the key vectors. The output of MHA$(\cdot)$ is obtained by concatenating the outputs of all heads and then applying a final linear transformation:

$$\mathrm{MHA}\left(X\right) = W_O\left(\mathrm{Cat}\left(Attn_1, \cdots, Attn_h\right)\right) \tag{13}$$

where $W_O$ is a learnable weight matrix. Feature maps processed by MHA$(\cdot)$ are concatenated along the channel dimension to form a new feature map. The formula is abbreviated as:

$$Y_{\mathrm{cat}} = \left[\mathrm{MHA}\left(Y_1\right), \mathrm{MHA}\left(Y_2\right), \mathrm{MHA}\left(Y_3\right)\right] \tag{14}$$

where $Y_1$, $Y_2$, and $Y_3$ are the three-layer feature maps. $[\cdot]$ represents the channel concatenation operation.

Subsequently, $Y_{\mathrm{cat}}$ is processed by the fusion module $G$ which utilizes a dynamic mechanism for fusion adjustment. Before being processed by $G$, $Y_{\mathrm{cat}}$ is further processed by the MHA mechanism. The fusion process applies different levels of attention to each feature channel, thereby distinguishing the importance of different features. The operation is represented as:

$$Y_{\mathrm{fused}} = G\left(\mathrm{MHA}\left(Y_{\mathrm{cat}}\right)\right) \tag{15}$$

where $G$ is the fusion function that helps us integrate and refine the multi-scale features. It integrates the information obtained from different scales and different convolutional layers, enhancing robustness and discriminability, and establishing a solid foundation for subsequent fault diagnosis.

### *3.3. Transformer encoder imbedded PE*

As artificial intelligence advances, the incorporation of the self-attention mechanism within the transformer framework eliminates the need for conventional recursive methods in the training process. It calculates the degree of correlation between each position of the input sequence and all other positions to understand the entire input sequence without considering the order of positions. However, relying solely on the self-attention mechanism disregards sequential operations due to parallel computation, leading to the loss of sequence information.

Aiming at the problem that the self-attention mechanism cannot directly learn position information, this study proposes a method of performing position-encoding operations on input sequences to integrate absolute positional data into the model. The position encoding is typically a fixed-dimension vector with the same dimension as the input embedding and is added to the input embedding. Before applying position encoding, residual blocks are introduced to preliminarily process the input sequences. The proposed method first processes input sequences through residual blocks and then applies position encoding, provides accurate positional information by fully extracting input sequence features. The PE method in this study utilizes sine and cosine functions of different frequencies, as described by the following formula: For position pos and dimension $i$, the calculation formula for the position encoding PE is:

$$\mathrm{PE}_{(\mathrm{pos},2i)} = \sin\left(\frac{\mathrm{pos}}{10000^{\frac{2i}{d_{model}}}}\right) \tag{16}$$

$$\mathrm{PE}_{(\mathrm{pos},2i+1)} = \cos\left(\frac{\mathrm{pos}}{10000^{\frac{2i}{d_{\mathrm{model}}}}}\right) \tag{17}$$

where pos is the position, ranging from 0 to the sequence length minus 1. $i$ is the dimension index, ranging from 0 to $d_{\mathrm{model}}/2-1$. $d_{\mathrm{model}}$ is the dimension of the model, that is, the dimension of the input embedding and the position encoding. The position encoding PE is added to the input embedding $E$ to obtain the vector $X$ that is finally input to the transformer layer:

$$X = E + \mathrm{PE} \tag{18}$$

The position encoding enables the transformer to effectively handle the sequential information in the input sequence without requiring an additional recurrent structure. It allows the transformer to accurately locate the positional relationship of fault-related information in the sequence, thereby enhancing the accuracy and comprehensiveness of fault feature extraction and providing a more reliable basis for subsequent fault diagnosis and analysis.

### *3.4. Joint feature extraction of MSRCNN and transformer*

In the field of DL, a network capable of simultaneously learning local and global features and effectively fusing them can achieve more accurate judgments and decisions when faced with classification or diagnosis tasks. In the proposed method, particular emphasis is placed on the synergetic effect of the MSRCNN branch and the transformer branch in feature extraction.

Firstly, MSRCNN's distinctive structural design enables efficient extraction of local characteristics of the input signal. Within its architecture, the multiple convolutional layers of diverse scales work in tandem, thereby enabling a meticulous and in-depth dissection of the signal's local particulars. Meanwhile, the transformer encoder leverages its powerful self-attention mechanism to extract global features.

After completing their respective feature extractions, the extracted features are concatenated, enabling the model to comprehend the input signal from multiple perspectives. Finally, the concatenated feature sequence is fed into the classifier, effectively accomplishing the fault diagnosis task of rolling bearings. The process is represented by the following formula:

$$E = \mathrm{Conv1}D\left(X\right) + \mathrm{Pos} \tag{19}$$

$$F = \mathrm{FCL}\left(\mathrm{Cat}\left(\mathrm{Trans\_Pos}\left(E\right), \mathrm{MSRCNN}\left(X\right)\right)\right) \tag{20}$$

$$Y = \mathrm{softmax}\left(F\right) \tag{21}$$

where $\mathrm{Pos} = [p_1, \cdots, p_T]$ represents the absolute position information injected into the encoder. Trans_Pos$(\cdot)$ represents feature extraction through the transformer encoder with PE. MSRCNN$(\cdot)$ represents feature extraction through via Multiscale residual convolutional. FCL$(\cdot)$ represents the FCL. softmax$(\cdot)$ is a function that normalizes the obtained final feature sequence with exponential values. $F$ and $Y$ refer to the output of the FCL and the output of the model, respectively.

### *3.5. TL*

Typically, model training requires a sufficient number of fault samples to achieve optimal accuracy. However, acquiring such samples in real-world production environments can be financially burdensome, making it difficult to obtain enough labeled data for training.To overcome this challenge, a TL strategy, based on model fine-tuning, is employed. The approach leverages a pre-trained model that has already learned robust feature representations from large-scale datasets. These learned features are then transferred to the current bearing fault diagnosis model, enabling it to adapt effectively to the target domain with limited labeled data. This allows the model to generalize well and achieve high accuracy even when only a small amount of labeled target domain data is available.

The model training process consists of two stages, starting with pre-training on the source domain dataset. In this phase, the entire network is trained without any layer restrictions, utilizing cross-entropy loss for optimization. Regularization techniques, such as dropout, are applied to prevent overfitting. This enables the model to learn rich feature representations that capture both local and global dependencies, thereby equipping it with generalizable features. Once this foundational training is complete, the model progresses to the fine-tuning phase, where it is adapted to the target domain. In this phase, the feature extraction layers, including the convolutional and transformer modules, are frozen to retain the learned weights, ensuring that the generalizable features are preserved. Only the final classification layers, comprising fully connected layers followed by a softmax output layer, are trained using the small labeled dataset from the target domain. This selective fine-tuning approach allows the model to maintain computational efficiency while mitigating the need for large volumes of labeled data, offering a practical solution for industrial fault diagnosis tasks where acquiring fault samples is often resource-intensive.

## 4. Experiments and results

In practical applications, the working environment of mechanical systems is often highly complex and highly variable. On one hand, given production requirements, working conditions frequently change, so it is unrealistic to collect and label enough training samples to ensure the classifier's robustness across working loads. Therefore, it is crucial that the feature extractor and classifier trained with samples from one working condition can classify samples in another working condition. On the other hand, in industrial production, obtaining training samples of different categories of bearings under similar working conditions is equally challenging. Hence, it is also of great significance to train the classifier with existing samples of faulty bearings to classify samples from other bearing categories. This section investigates the performance of the proposed model in these two scenarios. In order to demonstrate the superior performance of the MSCRT framework, a series of comparative experiments have been conducted. These include an evaluation of the effectiveness of TL under variable loads and an assessment of the model's performance with fine-tuning of TL across different bearing categories.

### *4.1. Data preprocessing and experimental setting*

To establish the groundwork for the experimental design, this study introduces the datasets. Subsequent sections will elaborate on the datasets, data processing methods, and the experimental design plan. Furthermore, the datasets are categorized into source domain and target domain datasets, aligning with the TL strategy. All experiments are conducted within a consistent operating environment, comprised of an Intel(R) Core(TM) i7-10 875 H CPU@2.30GHz, 32 GB of RAM, a NVIDIA GeForce RTX 2060 Laptop 6 GB graphics card, and the PyTorch deep learning framework.

*4.1.1. Dataset description.* In order to assess the efficacy and validity of the proposed diagnostic methodology, a series of empirical investigations have been conducted utilising two distinct data sets: the CWRU bearing dataset [38] and a Self-Collected bearing dataset. The following section will provide further insight into these two data sets.

**Table 1.** CWRU bearing dataset sample label.

| Label | Fault category | Fault size(mm) |
|---|---|---|
| 0 | Normal | — |
| 1 | B007 | 0.178 |
| 2 | B014 | 0.356 |
| 3 | B021 | 0.533 |
| 4 | IR007 | 0.178 |
| 5 | IR014 | 0.356 |
| 6 | IR021 | 0.533 |
| 7 | OR007 | 0.178 |
| 8 | OR014 | 0.356 |
| 9 | OR021 | 0.533 |

*4.1.1.1. CWRU bearing dataset.* The CWRU bearing dataset is based on tests on 6205-2RS SKF bearings. Vibration signals were collected from bearings under various working conditions at 12 kHz. The data were collected under four motor loads. Single-point faults of 0.178 mm, 0.356 mm, and 0.533 mm were introduced to the ball, inner raceway, and outer raceway. Nine categories exist for each working condition. The sample labels are in table 1.

*4.1.1.2. Self-Collected bearing dataset.* To further evaluate the performance of MSRCT, a more complex dataset containing compound faults was collected. The bearing type adopted in the Self-Collected dataset is 22 207CA/W33, and the sampling frequency is 96 kHz. The experiment included six bearing health states: normal, inner race fault, outer race fault, rolling element fault, inner race and rolling element compound fault, and outer race and rolling element compound fault. The test bench adopts the bearing type 22 207CA/W33. The testing environment of our dataset is illustrated in figure 4. Dataset sample labels are shown in table 2.

*4.1.2. Data preprocessing.* Each sample from the dataset is formed by selecting 2048 consecutive vibration data points from the original signal. Given the limited availability of training data, data augmentation strategies are crucial for significantly enhancing model performance. Consequently, the random sliding window sampling method has been adopted to mitigate the adverse effects of insufficient training data on the model, as illustrated in figure 5.

During the initial and final stages of signal acquisition, there will be unstable vibration signals. To avoid boundary effects influencing the experimental results, a boundary of length 1000 is removed at the start of the random sliding window sampling.The starting point for each sampling window is denoted by $X_j$ , where $j \in [1000, \text{length} - 1000 - 2048] \in \mathbb{N}$ . Here, 'length' refers to the total length of each vibration signal sequence in the dataset. The variable $\mathbb{N}$ represents the set of natural numbers.

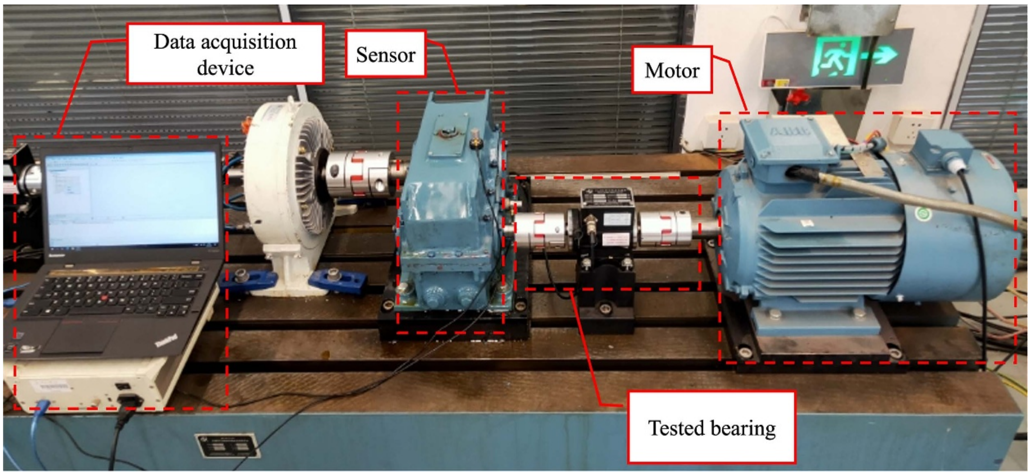


**Figure 4.** Experimental machine of Self-Collected two-stage gear drive test bench.

**Table 2.** Self-Collected bearing dataset sample label.

| Label | Fault category |
|---|---|
| 0 | Normal |
| 1 | Inner |
| 2 | Inner roller |
| 3 | Outer |
| 4 | Outer roller |
| 5 | Roller |

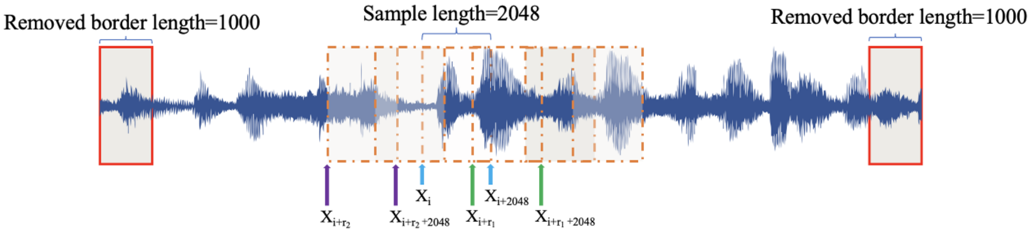


**Figure 5.** Random sliding window sampling.

It is evident that overlapping sampling frequently occurs during the application of the random sliding window for data augmentation, which may introduce information leakage across training, validation, and test sets. To mitigate this issue, a segmented sampling strategy is introduced to ensure strict data separation. Specifically, the original vibration signal is partitioned into non-overlapping segments, where each segment is exclusively assigned to the training, validation, or test set. Let $S_{\text{train}}$ , $S_{\text{val}}$ , and $S_{\text{test}}$ represent the respective data segments, satisfying the condition:

$$S_{\text{train}} \cap S_{\text{val}} = \emptyset, \quad S_{\text{train}} \cap S_{\text{test}} = \emptyset, \quad S_{\text{val}} \cap S_{\text{test}} = \emptyset \tag{22}$$

where each sample is drawn only from its designated segment to prevent cross-set contamination. Within each segment, the random sliding window sampling remains effective for data augmentation while ensuring that overlapping windows are confined within the same subset. The starting position for the window sampling, denoted as $X_i$, is randomly selected within the boundaries of its assigned segment. Specifically, the starting index $X_i$ is defined as:

$$X_i = X_{\text{start}} + i + r \tag{23}$$

where $X_{\text{start}}$ is the predefined starting index of a given segment, $L_s$ is the total length of the segment, and $r$ is a random offset coefficient. The index $i$ is a natural number chosen randomly from the range $[1000, L_s - 1000 - 2048]$, and $r$ is a random offset within the range of $[-50, 50]$ to add slight variations in the window positions. This ensures that all sampled windows remain strictly within their assigned subsets.

To further enhance data diversity while maintaining the integrity of the segmentation strategy, amplitude modulation and phase shift are applied exclusively within the training set:

$$X' = AX\cos(\theta) \tag{24}$$

**Table 3.** Hyperparameters of the MSRCT.

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| Transformer blocks | 6 | Embedding dimension | 96 |
| MLP ratio | 4 | Attention heads | 4 |
| Positional encoding dropout | 0.1 | QKV bias | True |
| Fusion dropout | 0.1 | Transformer dropout | 0.1 |
| CNN branch 1 dropout | 0.1 | CNN branch 2 dropout | 0.1 |
| CNN branch 3 dropout | 0.1 | Classifier dropout | 0.1 |
| Max epoch | 20 | Learning rate | 0.0001 |

where $A$ represents the modulation factor and $\theta$ denotes the phase shift. This targeted augmentation prevents information leakage while improving model robustness by exposing it to a broader range of variations within the training data.

By enforcing this segmented sampling approach, the proposed method ensures that data augmentation strategies effectively enhance model generalization while minimizing the risk of data leakage between training, validation, and test sets.

Furthermore, to efficiently alleviate the training complexity of the model, normalization is carried out on each sample prior to the commencement of training. This study adopts the $Z$-score normalization approach. Through this data preprocessing, all the data within a sample are transformed into a distribution characterized by a mean of 0 and a variance of 1. This method proves particularly effective for data distributions characterized by indistinct boundaries and potential extreme values. The normalization formula is specified as follows:

$$x^{*} = \frac{x-\mu}{\sigma} \tag{25}$$

where $x$ refers to the raw sample, $x^{*}$ represents the processed data, $\mu$ refers to the average value, and $\sigma$ represents the standard deviation of raw sample.

#### 4.1.3. Experimental setting.

This section presents multiple experiments based on the CWRU and Self-Collected bearing fault datasets, incorporating a TL strategy. These experiments encompass various bearing categories and operating conditions. To ensure the consistency of the experiments, the input sequence size is maintained consistently, with each sample comprising a fixed-length signal segment containing 2048 data points. Each fault category includes 2000 samples, which are divided into training, validation, and testing sets in a 7:1:2 ratio. The hyperparameters related to the experiment are shown in table 3. Among these, the learning rate and dropout rate play a crucial role in ensuring stable model performance. To facilitate stable convergence on a new model, a smaller learning rate was selected. Consistent with prior research in the field of fault diagnosis [39], a learning rate of 0.0001 was chosen for this study. As for the dropout rate, it typically ranges from 0.1 to 0.5, and in this experiment, a relatively lower dropout rate was applied. This choice helps retain more neurons during training, enabling the model to learn richer feature representations while effectively preventing overfitting, which is particularly advantageous for transfer tasks. Furthermore, to ensure the model's stability during training, these hyperparameters were fine-tuned through various experiments, optimizing the model's generalization ability and ultimately enhancing the accuracy and reliability of the predictions.

#### 4.1.4. TL strategy.

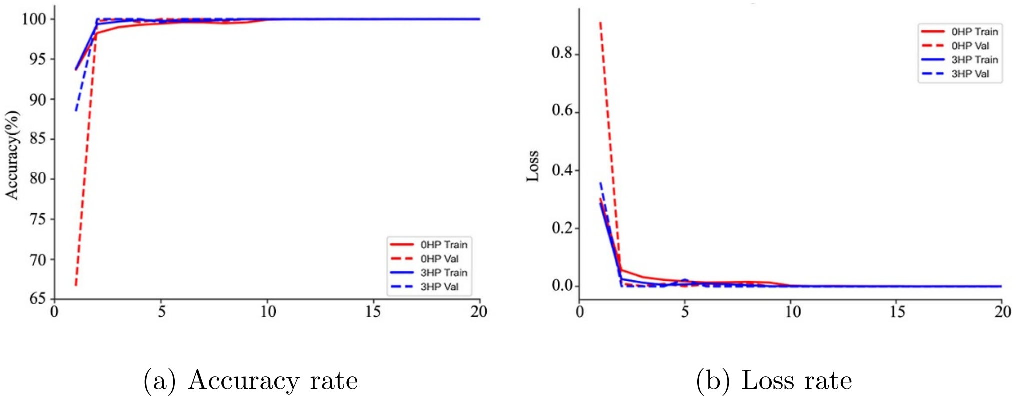
In the present study, experiments were conducted under diverse working conditions using the CWRU dataset, where one working condition serves as the source domain and the remaining conditions are used as target domains. Two simulation approaches were implemented to evaluate model performance: direct application and fine-tuning with a limited sample size. The direct application approach involved training the model on the source domain for 20 iterations, followed by validation using the target domain test set. In contrast, the fine-tuning approach involved training the model on the source domain while maintaining the learned weight parameters, with subsequent fine-tuning on the target domain using a small subset of samples. For both approaches, the target domain dataset was partitioned into training, validation, and testing sets in a ratio of 2:1:7. The primary objective of the experiments was to assess the accuracy of directly applying the weights from the source domain model to the target domain, and to compare it with the accuracy achieved after fine-tuning with the small sample size.

To ensure the robustness of the TL fine-tuning strategy, both the source and target domain datasets were designed with a balanced number of samples per fault category. Specifically, each fault category contained 2000 samples in both the source and target domains. This implies that, for the CWRU dataset, the training set comprised a total of 7000 samples, while the target domain dataset contained 2000 samples. Similarly, despite the Self-Collected dataset consisting of six distinct fault categories, each fault category was also allocated 2000 samples to maintain consistency. This approach ensures that the TL model operates under controlled sample conditions, enabling a fair comparison between the direct application and fine-tuning strategies.

### 4.2. Experimental results and analysis

#### 4.2.1. Fault diagnosis under different working conditions.

This section evaluates the performance of MSRCT on the bearing datasets across varying working conditions. The data of 0HP and 3HP from the CWRU dataset are selected, where HP denotes HP. Under the low load of 0HP, the vibration amplitude caused by bearing faults may be small, with relatively



(a) Accuracy rate

(b) Loss rate

**Figure 6.** The training accuracy and loss function of the CWRU dataset under the working conditions of 0HP and 3HP.

simple frequency components. In contrast, in the high load condition of 3HP, the vibration amplitude increases due to additional stress and friction, leading to the excitation of more complex frequency components. These differences provide a basis for testing the robustness of MSRCT. Additionally, to ensure a fair comparison, the training parameters across various working conditions remain consistent, encompassing input sample size, total number of samples, and learning rate.

In figure 6(a), the accuracy curve throughout the training process demonstrates a notable upward trend. In the initial stages, the accuracy is relatively low, indicating that the model is in the early stages of learning and understanding the data features. As the number of training epochs increases, the model gradually grasps the data features, and the accuracy steadily rises. The experimental results demonstrate that the model's learning capability steadily enhances, thereby progressively optimizing its predictive performance.

In figure 6(b), the entire loss function curve starts from a relatively high initial point and gradually approaches lower values in a smooth manner. This smooth decreasing trend indicates the stability of the training process, with no drastic fluctuations or abnormal jumps, indicating that the model optimizes its parameters systematically and gradually converges to a stable state. As the loss function curve continues to decrease, the corresponding training accuracy curve shows an upward trend, highlighting a close positive correlation between the reduction in loss values and the improvement in model prediction accuracy. As the model continuously reduces prediction errors, its classification or prediction capabilities for the data are also gradually enhanced, further validating the effectiveness of the training process and the improvement in model performance.

To exclude the extreme identification of individual fault categories, a confusion matrix was used to analyze the test set results. Figure 7(a) shows that in the CWRU dataset under the 0HP working condition, the test set accuracy reached 99.9%. Eighteen samples of fault category IR014 were misdiagnosed as IR021, and two samples of fault category IR007 was misdiagnosed as IR021. Figure 7(b) indicates that in the CWRU dataset under the 3HP working condition, the test set accuracy is almost the same as that under 0HP. Only two samples of fault category B021 were misdiagnosed as B007. The evidence suggests that the model effectively identifies rolling bearings under various fault conditions, maintaining high accuracy and stability across differing loads.

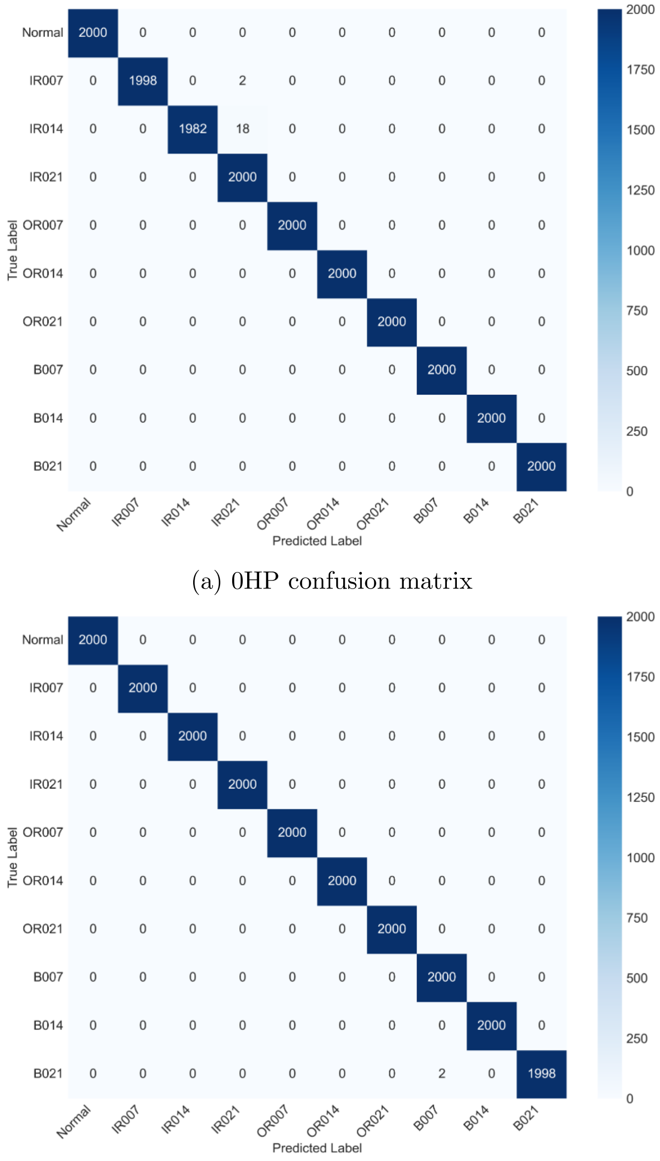
#### 4.2.2. Small sample TL diagnosis on the CWRU dataset.

This section assesses the TL performance of the model under varied working conditions. The emphasis is on determining if the model's prediction accuracy, post fine-tuning via TL between datasets with similar characteristics but varying HP, shows improvement when compared to its direct application on the target domain.

In real-world production processes, faulty bearing samples are often scarce and typically available only under specific working conditions. Acquiring faulty bearing samples under different working conditions is often expensive. To address the problem, a training strategy emerged: pre-training a source domain model and then fine-tuning it with a small sample size for the target domain. By leveraging a large dataset to train the source domain model, this strategy effectively tackles the issue of diagnosing faults in bearings with similar features across different operating conditions, even with limited samples. This approach holds significant practical significance.

In this part of the experiment, 12 transfer tasks were set up using the CWRU dataset. The source domain dataset was divided into training, validation, and test sets in a 7:1:2 ratio, while the target domain dataset was split in a 2:1:7 ratio. The experiment measured the accuracy of directly applying the weights saved from the source domain model to the target domain, as well as the test accuracy of the target domain model after fine-tuning with a small sample size.

In table 4, it can be observed that during the 12 transfer tasks, the direct application accuracy is relatively high for the transfer tasks between the 0HP and 1HP operating conditions, including both directions (0HP to 1HP and 1HP to 0HP). That may be attributed to the close proximity of bearing fault features resulting from the small difference in load. In contrast, for other tasks with varying working conditions, the test accuracy in the target domain remains high after applying the fine-tuning training strategy. The



(a) 0HP confusion matrix

(b) 3HP confusion matrix

**Figure 7.** The confusion matrix for the test sets under 0HP and 3HP operating conditions.

**Table 4.** TL performance of the model on the CWRU dataset.

| Source domain | Target domain | Direct Acc | Transfer Acc | Improvement |
|---|---|---|---|---|
| 0HP | 1HP | 99.95% | 100.00% | 0.05% |
| 0HP | 2HP | 90.35% | 100.00% | 9.65% |
| 0HP | 3HP | 80.72% | 100.00% | 19.28% |
| 1HP | 0HP | 98.88% | 99.78% | 0.90% |
| 1HP | 2HP | 90.35% | 100.00% | 9.65% |
| 1HP | 3HP | 80.47% | 99.95% | 19.48% |
| 2HP | 0HP | 89.35% | 99.95% | 10.60% |
| 2HP | 1HP | 90.15% | 99.95% | 9.80% |
| 2HP | 3HP | 93.30% | 100.00% | 6.70% |
| 3HP | 0HP | 79.40% | 99.55% | 20.15% |
| 3HP | 1HP | 70.17% | 98.95% | 28.78% |
| 3HP | 2HP | 80.75% | 99.62% | 18.87% |

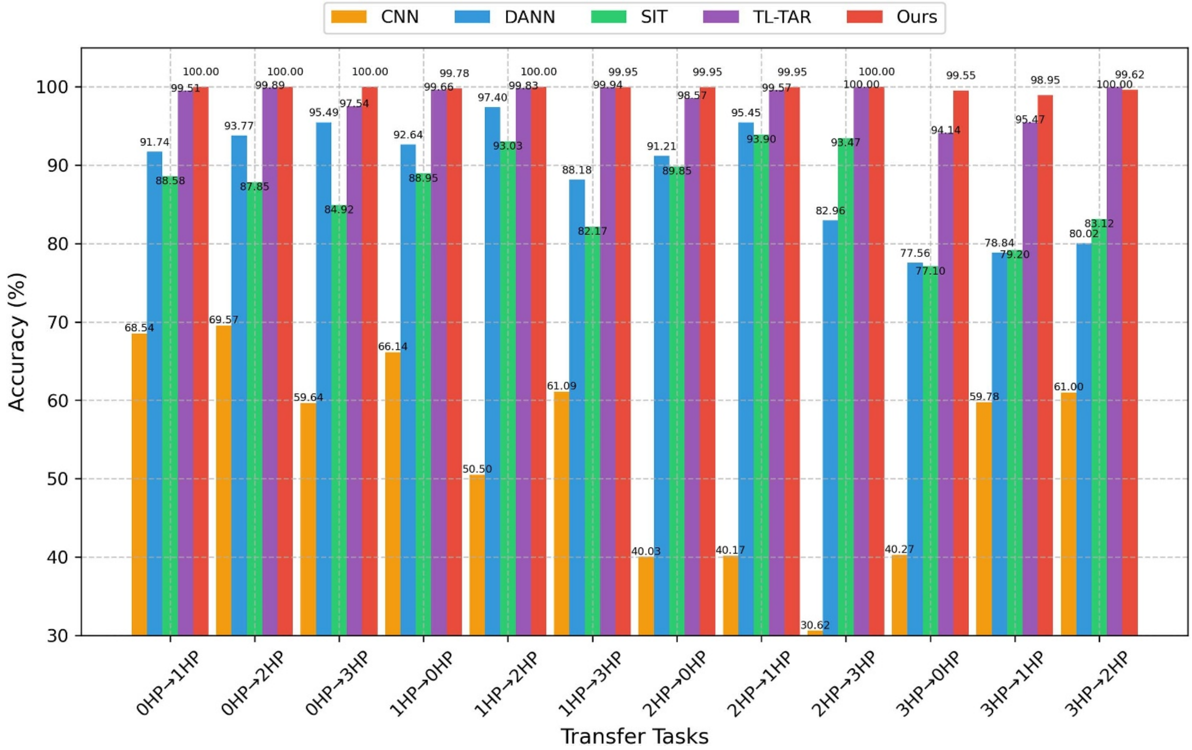


**Figure 8.** Accuracy comparison of different models on transfer tasks.

experimental results demonstrates the superior performance of MSRCT.

### 4.2.3. Comparative experiments on migration tasks of various models.

Advanced algorithms, including CNN [8], DANN [37], SIT [27], and TL-TAR [36], were employed for diagnostic performance comparisons on standardized task benchmarks. CNN is a highly representative model architecture in the field of DL and has been widely used and achieved remarkable results in numerous image and signal processing tasks. DANN focuses on addressing the problem of domain adaptation, and its unique architectural design can cope with the challenge of distribution differences when data come from different domains. In vibration signal analysis of fault diagnosis, after converting the vibration signal into a sequence form, the self-attention mechanism of SIT can comprehensively capture the complex correlations among various elements in the signal sequence, overcoming the limitations of traditional models in handling long-range dependencies. TL-TAR innovatively combines the advantages of TAR. Its dual-path feature extractor conducts feature extraction of the signal from different perspectives respectively. The transformer branch focuses on capturing the long-range dependencies of the signal and mining the correlation patterns of the signal on a global scale, while the ResNet branch effectively extracts local time-domain features with its one-dimensional convolution. Together, these branches comprehensively extract multi-scale signal features. In addition, its feature fusion module, classification head, and unique freezing and unfreezing mechanisms ensure strong adaptability in TL tasks.

(i) CNN effectively extract local feature patterns directly from the raw data.
(ii) The feature extractor, label classifier, and domain classifier of the DANN function collaboratively, facilitating domain adversarial training through the implementation of a gradient reversal layer. This architecture effectively learns domain-invariant features, thereby achieving cross-domain knowledge transfer and accurate diagnosis.
(iii) SIT comprehensively captures the complex correlation relationships among each element in the signal sequence, aiding the in-depth analysis of the differences of new models in mining the feature correlations of vibration signals using the attention mechanism.
(iv) Compared with TL-TAR, the performance of the new model in terms of multi-scale feature fusion strategy, small sample TL ability, and overall fault diagnosis performance can be evaluated in detail.

This section utilizes the CWRU dataset as the experimental data and conducts twelve transfer tasks. The same training samples are employed across all models, with multiple experiments conducted independently. It can be seen in figure 8 that among the 12 transfer tasks conducted, in terms of the average accuracy rate, the model we proposed is higher compared to other models. Although the TL-TAR model has a higher accuracy rate than our model in the task from 3HP to 2HP, in the 12 transfer tasks, the accuracy rate of our model remains at around 99% all the time, while the TL-TAR model is not as stable as ours. In the actual application process, the stability of fault diagnosis is particularly important. TL-TAR and the model we proposed perform well in the transfer tasks, it might be attributed to the fact that both models have designed similar freezing and unfreezing mechanisms of the feature extractor. DANN has a higher accuracy rate compared to the other two models. It arises from the introduction of the domain classifier mechanism, which better identifies the feature differences between the source and target domains.

This section selects the CWRU dataset as the experimental data and conducts 12 transfer tasks. Identical training samples are utilized across all models, and multiple experiments are conducted independently. Figure 8 shows that across the 12 transfer tasks, the proposed model achieves a higher average accuracy compared to other models. Although the TL-TAR model has a higher accuracy rate than our model in the task from 3HP to 2HP, in the 12 transfer tasks, the accuracy rate of our model remains at around 99% all the time, demonstrating greater stability compared to TL-TAR. In the actual application process, the stability of fault diagnosis is crucial. TL-TAR and the model we proposed perform well in the transfer tasks, it might be attributed to the fact that both models have designed similar freezing and unfreezing mechanisms of the feature extractor. DANN has a higher accuracy rate compared to the other two models. It introduces the domain classifier mechanism, which can better identify the feature differences between the source and target domains.

To further demonstrate and verify the ability of our model in classifying fault categories, this paper further studies the output features of these several different models. The high-dimensional features are reduced through *t*-SNE, then visualised. *t*-SNE assesses classification efficacy through data point distribution, inter-class distances, intra-class distances and so on. This study concentrates on inter-class and intra-class distances to assess the MSRCT model. Detailed definitions of these measures are as follows:

(i) Intra-class distance: It reflects the closeness of the data within the same class. A smaller intra-class distance implies that the data points of the same class are close to each other in the feature space, and the similarity of the data is high. Its performance enhances classification accuracy by reducing misclassification within the same class.

(ii) Inter-class distance: It embodies the degree of difference between data of different classes. A larger inter-class distance indicates that the data points of different classes are far apart in the feature space, and the discrimination between classes is high. It helps the classifier better distinguish different classes and enhance the reliability of classification.

Based on the *t*-SNE visualizations in figure 9, we can observe the feature clustering of various models, which highlights their ability to optimize both intra-class and inter-class distances. Figures 9(a)–(c) show that CNN, DANN, and SIT models exhibit significant overlap between different categories. This overlap indicates that the intra-class distances are relatively large in these models, which compromises their ability to distinctly separate fault types and leads to poor classification performance. The data points within the same class are not well-clustered, resulting in classification errors and reduced accuracy.

In contrast, figures 9(d) and (e) demonstrate that TL-TAR and MSRCT exhibit significantly better clustering, with the fault samples of the same class being tightly grouped. This reduction in intra-class distance indicates that both models effectively minimize within-class variance, contributing to better classification performance. However, MSRCT stands out by not only achieving well-separated intra-class clusters but also exhibiting significantly larger inter-class distances compared to TL-TAR. The greater inter-class distance in MSRCT enhances its ability to effectively distinguish between different fault types, leading to improved classification accuracy. While both TL-TAR and MSRCT exhibit small intra-class distances, MSRCT further excels by maximizing inter-class separation, resulting in superior class discrimination. This ability to tightly cluster intra-class data while increasing inter-class separation contributes to MSRCT's exceptional performance in bearing fault diagnosis, ensuring both high accuracy and reliable fault detection.

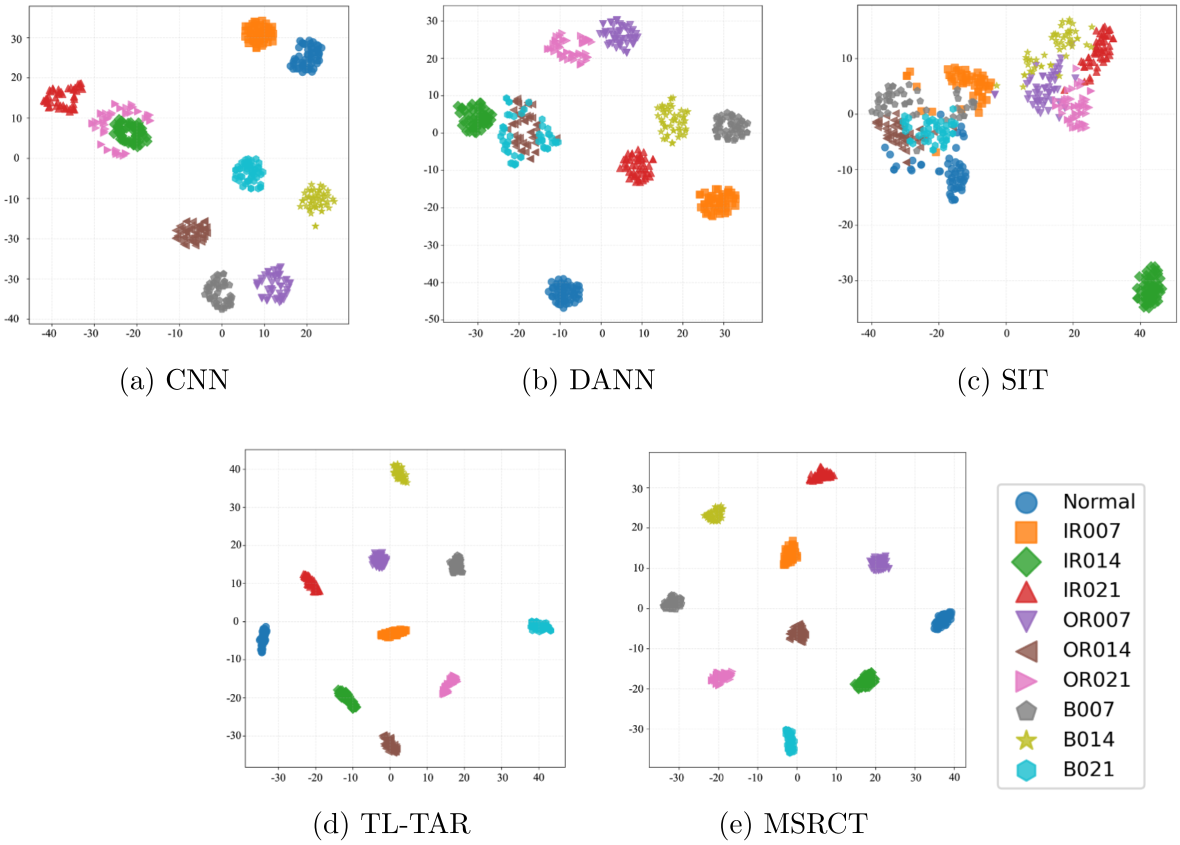
#### 4.2.4. Small sample TL diagnosis from CWRU dataset to self-collected dataset.

This section evaluates the feasibility of the TL fine-tuning strategy across different bearing categories.The CWRU dataset is the source domain, while a Self-Collected dataset is the target domain for small-sample fine-tuning. The source domain is partitioned into training, validation, and testing sets in a ratio of 7:1:2, while the target domain is organised in a 2:1:7 ratio, encompassing 2000 samples per fault category.

In the preliminary experiment, the model is pre-trained in the source domain and its classification layer adjusted to align with the target domain. The feature extraction layer's trainable parameters are frozen and the model fine-tuned on the target domain's training set, and the model's predictive accuracy is assessed on the test set.

In the second experiment, the model is trained on the Self-Collected fault dataset and evaluated. The dataset is divided into three sets: training, validation, and test. The ratio of data allocated to each set is 7:1:2. The model is trained and tested on the dataset to diagnose bearing faults.

Figure 10 exhibits the variation of test set accuracy in the training process. After just 2 training epochs, the model using the TL strategy with model fine-tuning attained an accuracy of approximately 99.3%. In comparison, the model directly trained on the target domain achieved over 99% accuracy in just five epochs. The TL strategy of model fine-tuning facilitates a transition from global to local training structures within the target domain dataset, thereby reducing training complexity. Moreover, despite disparities in features between the source and target domain datasets, their feature spaces exhibit similarities. By freezing the feature extraction layer, the model remains capable of comprehensively extracting features from



(a) CNN (b) DANN (c) SIT

(d) TL-TAR (e) MSRCT

**Figure 9.** The feature visualization of different models.

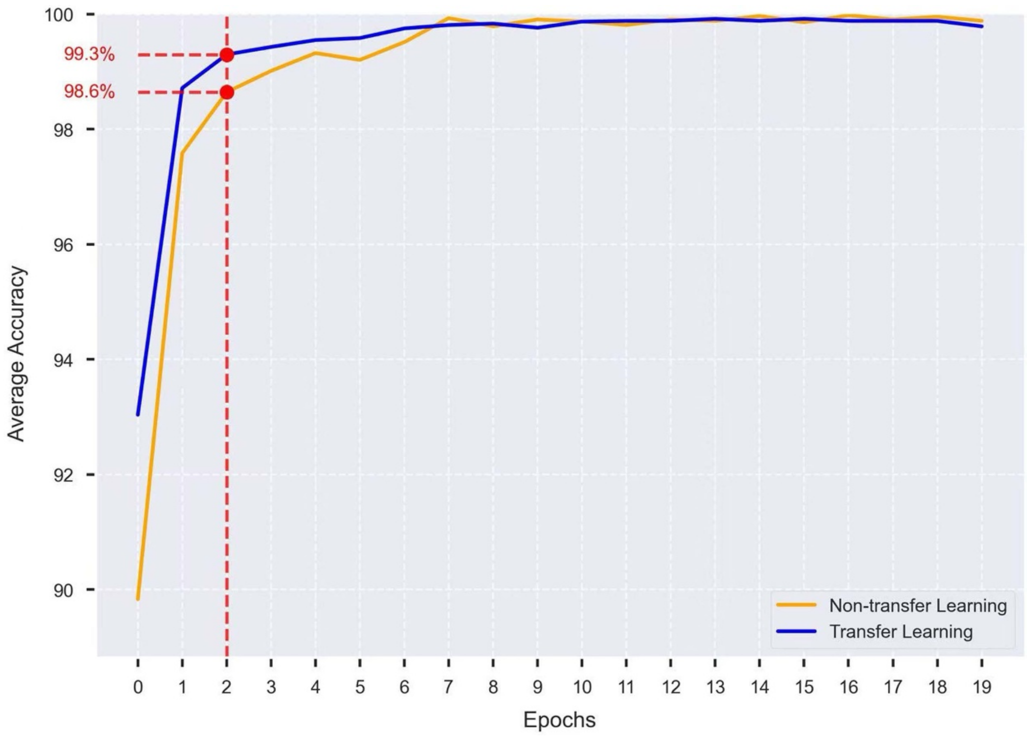


**Figure 10.** The process of model training iteration by TL and directly applied.

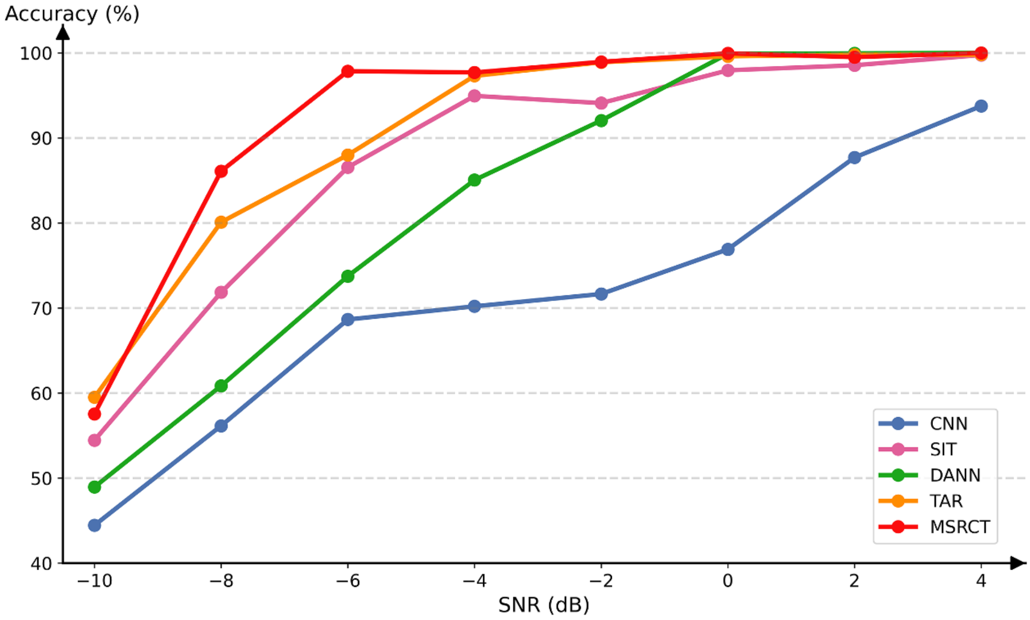


**Figure 11.** Comparison of accuracy on signals with different SNR.

the target domain's feature space. Overall, the integration of the proposed model with TL enables more effective and precise predictions, even with small sample sizes.

*4.2.5. Performance under noise environment.* In practical industrial environments, equipment operation is often accompanied by various types of noise, such as mechanical vibrations, electromagnetic interference, and environmental noise. These noises can interfere with or even drown out the fault characteristics, which directly affects the accuracy of fault diagnosis. Therefore, ensuring the effective performance of intelligent diagnostic models in noisy environments is of paramount importance. In this section, we validate the diagnostic accuracy of the proposed model under noisy conditions. Gaussian white noise of varying intensity is added to the raw signal to generate a noise dataset. A common method for evaluating the strength of signal noise is the SNR, defined as:

$$\mathrm{SNR_{dB}} = 10\log_{10}\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right) \tag{26}$$

where $P_{\mathrm{signal}}$ and $P_{\mathrm{noise}}$ represent the power of the signal and the noise, respectively.To better demonstrate the robustness of MSRCT in noisy environments, experiments were conducted on noise datasets with SNRs ranging from -10 dB to 4 dB. The results of MSRCT and other comparison methods are shown in figure 11.

From the figure 11, we observe that, under the influence of noise, TAR and MSRCT exhibit superior diagnostic performance compared to other methods. In particular, MSRCT demonstrates excellent performance across all SNRs. When the SNR is low, indicating significant noise impact, MSRCT's accuracy is much higher than that of other methods, showcasing its superior performance in noisy environments.The results indicate that the proposed MSRCT exhibits the strongest robustness and the highest diagnostic accuracy in noisy environments, making it more suitable for application in complex industrial settings.

*4.2.6. Ablation study of MSRCT.* In this section, we present an ablation study to evaluate the impact of different components in the MSRCT model. The experiments were conducted using the CWRU 3HP load dataset, with all experiments performed in identical experimental environments to ensure a fair comparison across different model configurations.

Three variations of the MSRCT model were tested: (1) MSRCT without the transformer, (2) MSRCT without MSRCNN, and (3) MSRCT without data augmentation. For the model without data augmentation, a fixed step-size window was employed for sampling, in contrast to the dynamic window approach used in the full model. The results of these experiments are summarized in table 5.

The full MSRCT model, incorporating all components, demonstrated the highest performance across all evaluation metrics. The incomplete MSRCT model, when excluding the transformer, the MSRCNN, or data augmentation, all exhibited a decline in performance compared to the complete MSRCT model, thereby underscoring the essential contribution of each component to the overall architecture.

The results emphasize the critical role of both the dual-branch structure and data augmentation in the MSRCT model. The dual-branch structure, combining MSRCNN and transformer, allows the model to effectively capture both local and global features, enhancing feature extraction. Meanwhile, the incorporation of data augmentation strategies, especially the random sliding window, significantly contributes to improving

**Table 5.** The ablation experiment of MSRCT.

| Models | Accuracy | Recall | F1-score |
|---|---|---|---|
| MSRCT without transformer | 0.9932 857 | 0.9933 169 | 0.9933 167 |
| MSRCT without MSRCNN | 0.9829 285 | 0.9830 219 | 0.9829 326 |
| MSRCT without data augmentation | 0.987 | 0.9879 069 | 0.9873 294 |
| MSRCT | **0.9992 142** | 0.9992 187 | 0.9992 191 |

*Note*: The bold entries indicate the best performance under the corresponding evaluation metrics.

**Table 6.** The computational complexity experiment of MSRCT.

| Models | Accuracy | Inference time(ms) | Model size(MB) | Total parameters |
|---|---|---|---|---|
| CNN(Re-Impl.) | 0.67 129 | 0.000 989 | 0.16 093 | 41 450 |
| DANN(Re-Impl.) | 0.91 007 | 0.02 394 | 16.54 393 | 4331 339 |
| SIT(Re-Impl.) | 0.92 283 | 0.1578 | 3.83 195 | 315 338 |
| TL-TAR(Re-Impl.) | 0.99 336 | 0.75 302 | 89.99 076 | 23 590 538 |
| MSRCT | **0.99 921** | 0.87 646 | 1.4499 | 344 591 |

*Note*: The bold entries indicate the best performance under the corresponding evaluation metrics.

the model's generalization capability by mitigating overfitting. This approach enables the model to better adapt to variations in the dataset, ensuring that it learns more robust features that are not overly dependent on specific data patterns. These findings demonstrate the importance of combining multiple components to enhance overall model performance, where each element contributes to better generalization and fault diagnosis accuracy under varying conditions.

*4.2.7. Computational complexity experiment of MSRCT.* To evaluate the practical applicability of the MSRCT model in real-world industrial settings, we conducted an experiment using the CWRU 3HP load dataset. The analysis of MSRCT's computational complexity, including accuracy, inference time, model size, and parameter efficiency, is summarized in table 6 and compared with other models.

Although MSRCT has a slightly longer inference time compared to models like TL-TAR, SIT, DANN, and CNN, this trade-off is justified by its superior accuracy. In real-time fault diagnosis, the slight increase in processing time is outweighed by the higher accuracy, which ensures more reliable fault detection and reduces maintenance errors. The model's higher accuracy is crucial in industrial environments where undetected faults can have severe consequences.

In terms of model size and parameters, MSRCT stands out for its smaller size compared to larger models like DANN and TL-TAR. Despite this compact design, it maintains high diagnostic accuracy, making it particularly suitable for deployment in resource-constrained environments, such as edge devices and embedded systems. Its reduced size and efficient use of parameters ensure that it can be integrated into existing systems with minimal hardware requirements, offering a clear advantage in practical applications.

In conclusion, MSRCT provides an optimal balance between accuracy, inference time, and parameter efficiency. Its smaller model size and lower computational demands, combined with high accuracy, make it an excellent choice for bearing fault diagnosis in resource-limited industrial environments.

## 5. Conclusion

To address the challenges of limited sample size and variability in working conditions for bearing fault diagnosis, this study proposes MSRCT, a dual-branch model, based on MSRCNN and transformer, designed to simultaneously extract global and local features of fault samples. In the transformer branch, a position encoding module is introduced to enhance the model's ability to perceive positional information, thereby improving classification accuracy. Additionally, residual blocks are embedded in the dual-branch structure to effectively mitigate gradient vanishing and explosion issues, supporting the training of deeper networks. To further optimize model performance, a TL strategy is employed, where the weights of the source domain model are transferred to the target domain through a freezing and unfreezing mechanism, enabling efficient fine-tuning with limited samples. Experiments on the CWRU and Self-Collected datasets demonstrate that the model outperforms existing models in terms of accuracy and stability under varying working conditions. Future work will focus on the development of intelligent diagnosis methods for larger datasets and some complex working conditions with uncertainties.

Looking ahead, there are several potential avenues for future research and improvements. First, when scaling applications to larger datasets, limitations in computational efficiency and storage capacity may arise. Providing distributed computing and cloud computing will enable efficient data processing. Furthermore, it would be valuable to explore the application of the MSRCT to complex and dynamic industrial scenarios. Integrating meta-learning [40] into MSRCT is a promising solution to enhance adaptability to unknown conditions. Meta-learning leverages existing knowledge and experience to enable rapid learning, allowing intelligent models to quickly adapt to new tasks and thereby enhancing their generalization capability. This will be a key focus of future research.

## Data availability statement

The data cannot be made publicly available upon publication because no suitable repository exists for hosting data in this field of study. The data that support the findings of this study are available upon reasonable request from the authors.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 52275101.

## ORCID iDs

Shuzhen Han https://orcid.org/0009-0004-1586-5354
Jianfei Li https://orcid.org/0009-0003-4414-097X
Ke Pang https://orcid.org/0009-0006-4227-9558
Dong Zhen https://orcid.org/0000-0002-5047-3346
Guojin Feng https://orcid.org/0000-0001-9937-910X
Fujun Tian https://orcid.org/0009-0009-2818-1425
Pingjuan Niu https://orcid.org/0000-0001-7888-7690

## References

[1] Yang L, Yang Z, Song S, Li F and Chen C L P 2023 Twin broad learning system for fault diagnosis of rotating machinery *IEEE Trans. Instrum. Meas.* **72** 1–12

[2] Zhu Z, Lei Y, Qi G, Chai Y, Mazur N, An Y and Huang X 2023 A review of the application of deep learning in intelligent fault diagnosis of rotating machinery *Measurement* **206** 112346

[3] Pu H, Zhang K and An Y 2023 Restricted sparse networks for rolling bearing fault diagnosis *IEEE Trans. Ind. Inform.* **19** 11139–49

[4] Riaz S, Elahi H, Javaid K and Shahzad T 2017 Vibration feature extraction and analysis for fault diagnosis of rotating machinery-a literature survey *Asia Pac. J. Multidiscip. Res.* **5** 103–10

[5] Zhao X and Guo H 2023 Rolling bearing fault diagnosis model based on DSCB-NFAM *Meas. Sci. Technol.* **35** 015029

[6] Chen X, Yang R, Xue Y, Huang M, Ferrero R and Wang Z 2023 Deep transfer learning for bearing fault diagnosis: a systematic review since 2016 *IEEE Trans. Instrum. Meas.* **72** 1–21

[7] Hoang D T and Kang H J 2019 A survey on Deep Learning based bearing fault diagnosis *Neurocomputing* **335** 327–35

[8] Kiranyaz S, Avci O, Abdeljaber O, Ince T, Gabbouj M and Inman D J 2021 1D convolutional neural networks and applications: a survey *Mech. Syst. Signal Process.* **151** 107398

[9] Janssens O, Slavkovikj V, Vervisch B, Stockman K, Loccufier M, Verstockt S, Van de Walle R and Van Hoecke S 2016 Convolutional neural network based fault detection for rotating machinery *J. Sound Vib.* **377** 331–45

[10] Lu C, Wang Z and Zhou B 2017 Intelligent fault diagnosis of rolling bearing using hierarchical convolutional network based health state classification *Adv. Eng. Inf.* **32** 139–51

[11] Liu R, Meng G, Yang B, Sun C and Chen X 2017 Dislocated time series convolutional neural architecture: an intelligent fault diagnosis approach for electric machine *IEEE Trans. Ind. Inf.* **13** 1310–20

[12] Wen L, Li X, Gao L and Zhang Y 2018 A new convolutional neural network-based data-driven fault diagnosis method *IEEE Trans. Ind. Electron.* **65** 5990–8

[13] Yao Y, Wang H, Li S, Liu Z, Gui G, Dan Y and Hu J 2018 End-To-End convolutional neural network model for gear fault diagnosis based on sound signals *Appl. Sci.* **8** 1584

[14] Youcef Khodja A, Guersi N, Saadi M N and Boutasseta N 2020 Rolling element bearing fault diagnosis for rotating machinery using vibration spectrum imaging and convolutional neural networks *Int. J. Adv. Manuf. Technol.* **106** 1737–51

[15] Pan J, Qu L and Peng K 2021 Sensor and actuator fault diagnosis for robot joint based on deep CNN *Entropy* **23** 751

[16] Zhang W, Peng G, Li C, Chen Y and Zhang Z 2017 A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals *Sensors* **17** 425

[17] 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 *Mech. Syst. Signal Process.* **100** 439–53

[18] Sonmez E, Kacar S and Uzun S 2023 A new deep learning model combining CNN for engine fault diagnosis *J. Braz. Soc. Mech. Sci. Eng.* **45** 644

[19] Jiao J, Zhao M, Lin J and Liang K 2020 Residual joint adaptation adversarial network for intelligent transfer fault diagnosis *Mech. Syst. Signal Process.* **145** 106962

[20] Zhang W, Li X and Ding Q 2019 Deep residual learning-based fault diagnosis method for rotating machinery *ISA Trans.* **95** 295–305

[21] Yan J, Kan J and Luo H 2022 Rolling bearing fault diagnosis based on markov transition field and residual network *Sensors* **22** 3936

[22] Yu S, Wang M, Pang S, Song L and Qiao S 2022 Intelligent fault diagnosis and visual interpretability of rotating machinery based on residual neural network *Measurement* **196** 111228

[23] Hu J, Li W, Zheng X, Tian Z and Zhang Y 2023 Prior knowledge-based residuals shrinkage prototype networks for cross-domain fault diagnosis *Meas. Sci. Technol.* **34** 105011

[24] Wang R, Dong E, Cheng Z, Liu Z and Jia X 2024 Transformer-based intelligent fault diagnosis methods of mechanical equipment: a survey *Open Phys.* **22** 20240015

[25] Lv H, Chen J, Pan T, Zhang T, Feng Y and Liu S 2022 Attention mechanism in intelligent fault diagnosis of

machinery: a review of technique and application *Measurement* **199** 111594

[26] Ahmad I, Ul Islam Z, Riaz S and Xue F 2024 MANS-Net: multiple attention-based nuclei segmentation in multi organ digital cancer histopathology images *IEEE Access* **12** 173530–9

[27] Yang Z, Cen J, Liu X, Xiong J and Chen H 2022 Research on bearing fault diagnosis method based on transformer neural network *Meas. Sci. Technol.* **33** 085111

[28] Tang X, Xu Z and Wang Z 2022 A novel fault diagnosis method of rolling bearing based on integrated vision transformer model *Sensors* **22** 3878

[29] Liu W, Zhang Z, Zhang J, Huang H, Zhang G and Peng M 2023 A novel fault diagnosis method of rolling bearings combining convolutional neural network and transformer *Electronics* **12** 1838

[30] Jiao Z, Pan L, Fan W, Xu Z and Chen C 2022 Partly interpretable transformer through binary arborescent filter for intelligent bearing fault diagnosis *Measurement* **203** 111950

[31] Pei X, Zheng X and Wu J 2021 Rotating machinery fault diagnosis through a transformer convolution network subjected to transfer learning *IEEE Trans. Instrum. Meas.* **70** 1–11

[32] He Q, Li S, Bai Q, Zhang A, Yang J and Shen M 2022 A siamese vision transformer for bearings fault diagnosis *Micromachines* **13** 1656

[33] Li X, Yuan P, Wang X, Li D, Xie Z and Kong X 2023 An unsupervised transfer learning bearing fault diagnosis method based on depthwise separable convolution *Meas. Sci. Technol.* **34** 095401

[34] Li X, Su K, Li D, He Q, Xie Z and Kong X 2024 Transfer learning for bearing fault diagnosis: adaptive batch normalization and combined optimization method *Meas. Sci. Technol.* **35** 046106

[35] Ai T, Liu Z, Zhang J, Liu H, Jin Y and Zuo M 2023 Fully simulated-data-driven transfer-learning method for rolling-bearing-fault diagnosis *IEEE Trans. Instrum. Meas.* **72** 1–11

[36] Hou S, Lian A and Chu Y 2023 Bearing fault diagnosis method using the joint feature extraction of Transformer and ResNet *Meas. Sci. Technol.* **34** 075108

[37] Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, March M and Lempitsky V 2016 Domain-Adversarial training of neural networks *J. Mach. Learn. Res.* **17** 1–35

[38] Smith W A and Randall R B 2015 Rolling element bearing diagnostics using the Case Western Reserve University data: a benchmark study *Mech. Syst. Signal Process.* **64-65** 100–31

[39] Han S, Sun S, Zhao Z, Luan Z and Niu P 2024 Deep residual multiscale convolutional neural network with attention mechanism for bearing fault diagnosis under strong noise environment *IEEE Sens. J.* **24** 9073–81

[40] Lei T, Hu J and Riaz S 2023 An innovative approach based on meta-learning for real-time modal fault diagnosis with small sample learning *Front. Phys.* **11** 1207381