兰州理工大学

科研成果汇总

学　　号：	231081101004
研 究 生：	张亚洲
导　　师：	赵小强 教授
研究方向：	故障诊断与寿命预测
论文题目：	面向旋转机械的深度学习故障诊断与寿命预测方法研究
学　　科：	控制理论与控制工程
学　　院：	自动化与电气工程学院
入学时间：	2023 年 9 月

目录

图书馆 文献检索报告

兰州理工大学图书馆信息咨询与学科服务部 报告编号：**R2026-0335**

机构：兰州理工大学 电气工程与信息工程学院

姓名：张亚洲 **[231081101004]**

著者要求对其在国内外学术出版物所发表的科技论著被以下数据库收录情况进行查证。

检索范围：

- 科学引文索引（Science Citation Index Expanded）： 1900年-2026年
- 工程索引（Engineering Index）： 1884年-2026年

检索结果：

检索类型	数据库	年份范围	总篇数	第一作者
SCI-E 收录	SCI-EXPANDED	2024 - 2026	7	7
EI 收录	EI-Compendex	2024	1	1

委托人声明：

本人委托兰州理工大学图书馆信息咨询与学科服务部查询论著被指定检索工具收录情况，经核对检索结果，附件中所列文献均为本人论著，特此声明。

作　　者（签字）：张亚洲

完 成 人（签字）：陈　超

完　成　日　期　：2026年4月23日

完成单位（盖章）：兰州理工大学图书馆信息咨询与学科服务部

（本检索报告仅供校内使用）

图书馆

文献检索报告
SCI-E 收录

兰州理工大学图书馆信息咨询与学科服务部

报告编号：R2026-0335 SCI-E 收录

数据库：科学引文索引 (Science Citation Index Expanded) 时间范围：2024年至2026年	作者姓名：张亚洲 作者单位：兰州理工大学 电气工程与信息工程学院	检索人员：陈　超 检索日期：2026年4月23日

检索结果：被 SCI-E 收录文献 7 篇

#	作者	地址	标题	来源出版物	文献类型	入藏号
1	**Zhang, YZ; Zhao, XQ; Peng, ZR; Hui, YY; Xu, RR**	[Zhang, Yazhou; Zhao, Xiaoqiang; Peng, Zhenrui; Hui, Yongyong; Xu, Rongrong] Lanzhou Univ Technol, Coll Elect & Informat Engn, Lanzhou 730050, Peoples R China.; [Zhao, Xiaoqiang; Peng, Zhenrui; Hui, Yongyong] Gansu Key Lab Adv Control Ind Proc, Lanzhou 730050, Peoples R China.	FCTransformer: An intelligent fault diagnosis method for multi-sensor information fusion based on Fourier convolution and transformer	*MECHANICAL SYSTEMS AND SIGNAL PROCESSING* 2026, 249: 114061.	J Article	WOS:001705848500001
2	**Zhang, YZ; Zhao, XQ; Peng, ZR; Xu, RR; Chen, P**	[Zhang, Yazhou; Zhao, Xiaoqiang; Peng, Zhenrui; Xu, Rongrong] Lanzhou Univ Technol, Coll Elect & Informat Engn, Lanzhou 730050, Peoples R China.; [Zhao, Xiaoqiang; Peng, Zhenrui] Gansu Key Lab Adv Control Ind Proc, Lanzhou 730050, Peoples R China.; [Chen, Peng] Lanzhou Petrochem Univ Technol, Coll Elect & Elect Engn, Lanzhou 730050, Peoples R China.	WD-KANTF: An interpretable intelligent fault diagnosis framework for rotating machinery under noise environments and small sample conditions	*ADVANCED ENGINEERING INFORMATICS* 2025, 66: 103452.	J Article	WOS:001490156300002
3	**Zhang, YZ; Zhao, XQ; Xu, RR**	[Zhang, Yazhou; Zhao, Xiaoqiang; Xu, Rongrong] Lanzhou Univ Technol, Sch Elect Engn & Informat Engn, Lanzhou 730050, Gansu, Peoples R China.	Feature and Joint Distribution Migration Alignment Method for Cross-Domain Fault Diagnosis of Rotating Machinery	*IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT* 2025, 74: 3525115.	J Article	WOS:001459634000031
4	**Zhang, YZ; Zhao, XQ; Peng, ZR; Xu, RR; Hui, YY**	[Zhang, Yazhou; Zhao, Xiaoqiang; Peng, Zhenrui; Xu, Rongrong; Hui, Yongyong] Lanzhou Univ Technol, Coll Elect & Informat Engn, Lanzhou 730050, Peoples R China.; [Zhao, Xiaoqiang; Peng, Zhenrui; Hui, Yongyong] Gansu Key Lab Adv	Cross-domain remaining useful life prediction for rolling bearings based on wavelet decomposition and dynamic calibrated domain adaptive networks	*MEASUREMENT* 2025, 251: 117278.	J Article	WOS:001451537000001

		Control Ind Proc, Lanzhou 730050, Peoples R China.				
5	**Zhang, YZ**; Zhao, XQ; Peng, ZR; Hui, YY; Xu, RR; Chen, P	[Zhang, Yazhou; Zhao, Xiaoqiang; Peng, Zhenrui; Hui, Yongyong; Xu, Rongrong] Lanzhou Univ Technol, Coll Elect & Informat Engn, Lanzhou 730050, Peoples R China.; [Zhao, Xiaoqiang; Peng, Zhenrui; Hui, Yongyong] Gansu Key Lab Adv Control Ind Proc, Lanzhou 730050, Peoples R China.; [Chen, Peng] Lanzhou Petrochem Univ Technol, Coll Elect & Elect Engn, Lanzhou 730050, Peoples R China.	An interpretable frequency-enhanced domain adaptive network for cross-domain fault diagnosis of rotating machinery	*APPLIED ACOUSTICS* 2025, 240: 110934.	J Article	WOS:001534922900001
6	**Zhang, YZ**; Zhao, XQ; Liang, HP; Chen, P	[Zhang, Yazhou; Zhao, Xiaoqiang] Lanzhou Univ Technol, Coll Elect & Informat Engn, Lanzhou 730050, Peoples R China.; [Liang, Haopeng] Lanzhou Univ Technol, Coll Comp & Commun, Lanzhou 730050, Peoples R China.; [Chen, Peng] Lanzhou Petrochem Univ Vocat Technol, Coll Elect & Elect Engn, Lanzhou 730060, Gansu, Peoples R China.	Multiscale dilated convolution and swin-transformer for small sample gearbox fault diagnosis	*APPLIED INTELLIGENCE* 2024, 54 (17-18): 7716-7732.	J Article	WOS:001246529500003
7	**Zhang, YZ**; Zhao, XQ; Xu, RR; Peng, ZR	[Zhang, Yazhou; Zhao, Xiaoqiang; Xu, Rongrong; Peng, Zhenrui] Lanzhou Univ Technol, Coll Elect & Informat Engn, Lanzhou 730050, Peoples R China.; [Zhang, Yazhou; Zhao, Xiaoqiang; Xu, Rongrong; Peng, Zhenrui] Gansu Key Lab Adv Control Ind Proc, Lanzhou 730050, Peoples R China.	A dual-stream temporal convolutional network for remaining useful life prediction of rolling bearings	*MEASUREMENT SCIENCE AND TECHNOLOGY* 2025, 36 (1): 016206.	J Article	WOS:001341515100001
					合计	7

图书馆

文献检索报告
EI 收录

兰州理工大学图书馆信息咨询与学科服务部

报告编号：R2026-0335 EI 收录

数据库：工程索引 (Engineering Index) 时间范围：2024年	作者姓名：张亚洲 作者单位：兰州理工大学 电气工程与信息工程学院	检索人员：陈　超 检索日期：2026年4月23日

检索结果：被 EI 收录文献 1 篇

#	作者	地址	标题	来源出版物	文献类型	入藏号
1	**Zhang, Ya-Zhou**; Zhao, Xiao-Qiang; Hui, Yong-Yong; Chen, Peng	College of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou; Key Laboratory of Advanced Control of Industrial Processes in Gansu Province, Lanzhou; College of Electrical and Electronic Engineering, Lanzhou Petrochemical University of Technology, Lanzhou	SA-DACNN gearbox fault diagnosis method based on multi-sensor data fusion 基于多传感器数据融合的SA-DACNN齿轮箱故障诊断方法	***Kongzhi yu Juece/Control and Decision*** 2024, 39 (11): 3699-3708.	Journal article (JA)	20244017127868
					合计	1

Mechanical Systems and Signal Processing 249 (2026) 114061

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

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

Full Length Article

FCTransformer: An intelligent fault diagnosis method for multi-sensor information fusion based on Fourier convolution and transformer

Yazhou Zhang [a], Xiaoqiang Zhao [a,b,*], Zhenrui Peng [a,b], Yongyong Hui [a,b], Rongrong Xu [a]

[a] College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China
[b] Gansu Key Laboratory of Advanced Control of Industrial Processes, Lanzhou 730050, China

ARTICLE INFO

Communicated by Olga Fink

Keywords:
Fault diagnosis
Rotating machinery
Information fusion
Fourier convolution
Transformer

ABSTRACT

Aiming at the problems of missing information and insufficient modelling in the frequency domain and time domain for the existing multi-sensor data fusion methods, an intelligent fault diagnosis method based on Fourier convolution and Transformer (FCTransformer) is proposed. Firstly, the method uses principal component analysis (PCA) to downscale high-dimensional and multi-modal sensor data, and converts them into RGB images by continuous wavelet transform (CWT). Then, a Fourier convolution (FC) is designed, which can not only effectively extract local periodic frequency domain features, but also enhances the nonlinear feature extraction capability of the model. Finally, a variational auto-encoding transformer (VAETransformer) is designed to achieve global dependency modelling, which enhances the identification for multi-sensor coupled features. To verify the effectiveness of the proposed method, experimental validation is carried out on the bearing and gearbox datasets. The results show that the average fault diagnosis accuracy of FCTransformer is 99.58% under normal sample conditions, and its average fault diagnosis accuracy is 96.10% under small sample conditions. This indicates that FCTransformer has superior fault diagnosis performance and can meet the requirements of rotating machinery fault diagnosis tasks.

1. Introduction

As key components of wind turbine generators, rail transport systems and aircraft engines, intelligent monitoring for rotating machinery has been an increasing hotspot [1,2]. However, rotating machinery inevitably suffers from failures such as wear, loosening and fatigue during high intensity and long working hours. In addition, the highly integrated character of rotating machinery makes it difficult to detect faults from within the machinery. Therefore, it is of practical significance to develop an efficient and accurate intelligent diagnosis method for rotating machinery to ensure the safe and stable operation [3–5].

Rotating machinery fault diagnosis methods are divided into model-based fault diagnosis methods and data-driven fault diagnosis methods [6]. Model-based fault diagnosis methods can only satisfy simple mechanical equipment fault diagnosis. However, in the face of integrated large-scale rotating machinery, it is difficult to achieve effective fault diagnosis. In recent years, with the arrival of sensor

* Corresponding author at: College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China.
E-mail address: xqzhao@lut.edu.cn (X. Zhao).

https://doi.org/10.1016/j.ymssp.2026.114061
Received 26 June 2025; Received in revised form 25 November 2025; Accepted 21 February 2026

technology and the 'big data' era, data-driven fault diagnosis methods for rotating machinery have become the research hotspot [7,8]. Among them, convolutional neural network (CNN) [9,10], recurrent neural network (RNN) [11,12] and their variants are widely used in the field of rotating machinery fault diagnosis, and have significant advantages in feature extraction and fault identification. For example, He et al [13] proposed a physically informative wavelet domain adaptive network for rotating machinery transfer learning. The network integrated interpretable wavelet knowledge into the convolutional layer, which can capture more domain-invariant features and increase the interpretability of the model. Zhao et al [14] proposed a multi-scale residual shrinkage network, which constructed a multi-scale residual shrinkage module to mine multi-scale features for the input data. Lu et al [15] designed a single source domain generalised fault diagnosis framework that utilised a priori knowledge embedded in the convolutional autoencoder for feature extraction. Since CNNs only rely on local receptive fields to extract features, they lack the abilities to capture global features and model long-range dependencies. To address these limitations, Transformers utilize self-attention mechanisms to compute the relationships between different positions of the input data, making them widely applicable for modeling spatiotemporal dependencies. For example, Zhang et al [16] constructed a multi-scale dilated convolution and swin-transformer gearbox fault diagnosis method. Xiao et al [17] designed a bayesian transformer for rotating machinery fault diagnosis, which can obtain the attention weights of random variables through model training.

The above methods offer new insights for transfer learning and noise-resistant diagnostics in rotating machinery, and have achieved encouraging results in experimental validation. However, there are still two limitations. (1) The computational complexity of Transformer increases quadratically with the input sequence length. In small sample fault diagnosis tasks, the model parameters may overfit due to insufficient training data. (2) The methods rely solely on vibration signals from a single sensor to extract input features. In rotating machinery operating environments, external factors can easily interfere with vibration signals from a single sensor, preventing the model from accurately extracting fault information. Therefore, to address the insufficient fault information representation capability of single sensor, researchers have begun exploring multi-sensor data fusion techniques.

Currently, multi-sensor data fusion technology is primarily categorized into multi-source information fusion at the data level, feature level, and decision level [18,19]. Feature level fusion relies on a model's feature learning and representation capabilities and requires hierarchical feature fusion. For example, Yang et al. [20] proposed a multi-sensor local and global feature fusion method that extracted local features through convolution and employed sparse Transformers for global feature perception learning. However, directly using convolution for feature extraction generates redundant information. Sun et al. [21] proposed a physics-knowledge-driven feature fusion and reconstruction network that achieved feature fusion of signals from different sensors through a dual-view integrated wavelet feature fusion module. However, the selection of wavelet kernels in the dual-view integrated wavelet feature fusion module relies on manual experience. Dong et al. [22] designed a synthetic cliff entropy to fuse signals from different sensors, and used a dual-regularized contrastive transformer for feature learning. However, the synthetic cliff entropy is only effective for the data collected by acceleration sensors. Decision level fusion utilizes submodels for feature extraction and fault determination, then fuses the outputs of these submodels to enhance the robustness of overall decisions. For example, Shao et al. [23] proposed a collaborative fault diagnosis method based on multi-sensor fusion, which employed stacked wavelet autoencoders for feature extraction and implemented multi-sensor feature fusion through an enhanced voting fusion strategy. However, stacked wavelet autoencoders can only extract local feature information and lack modeling of global dependencies. Gong et al. [24] proposed a fast fault diagnosis method based on improved CNN for multi-sensor data fusion, which fused fault features from different sensors in the model's fully connected layer. However, since fusion occurs at the end of the decision process, the long path leads to the lack of the shallow layer information. Data level fusion refers to the direct fusion of multi-sensor data through fusion strategies to maximize the retained feature information. For example, Lin et al. [25] utilized principal component analysis (PCA) and image transformation techniques to fuse multi-source sensor signals, and used parameter free attention mechanisms to adaptively extract domain-invariant features from input samples. However, image transformation techniques lack physical interpretation, which lead to unreliable diagnostic results. Ye et al. [26] proposed a deep integrated network based on multi-sensor information fusion for bearing fault diagnosis, which employed a composite exponential weighting fusion strategy to fuse multi-sensor information and utilized a cross-scale attention model to extract fault features. However, this approach demands high compatibility and dimensional consistency of input data and lacks error correction capabilities.

In summary, the above multi-sensor fusion methods lack the unified frequency-domain and time-domain modeling mechanisms, resulting in the failure to balance global feature capture with local detail preservation. Furthermore, when fusing high-dimensional and multi-modal data, incorrect fusion location selection leads to information redundancy. To address these issues, this paper proposes an intelligent fault diagnosis method based on FCTransformer for multi-sensor information fusion, which combines Fourier convolution with VAETransformer for the first time. FCTransformer primarily consists of a multi-sensor data fusion strategy, local feature extraction, and global dependency modeling. Firstly, the multi-sensor data fusion strategy is utilized to fuse multi-sensor data and generate an input dataset, this dataset provides the model with fault information from different locations. Secondly, Fourier convolution is employed for local feature extraction, which extracts local periodic fault features from the input dataset, enhancing the model's feature representation capability. Finally, global dependency modeling is performed using VAETransformer, which possesses regularization capabilities that prevent model overfitting and improve generalization performance. The contributions of this paper are summarised in the following four aspects:

(1) In order to reduce the missing of fault information and reveal the coupling relationship between different sensors, a data-level information fusion strategy is designed. The high-dimensional and multi-modal sensor data are downscaled by PCA, and RGB fused images are generated using CWT.

(2) In order to overcome the limitation of traditional convolution in dealing with frequency domain features, FC is designed for local feature extraction. FC not only enhances the nonlinear capability of the model, but also effectively mines the periodic frequency fault features.
(3) In order to prevent the model overfitting and eliminate redundant information, VAETransformer is developed for global dependency modelling. VAETransformer can capture long-range dependencies between different sensors and extracts global fault features.
(4) Experimental validation is performed on bearing and gearbox datasets. The results show that FCTransformer has excellent fault diagnosis performance compared to information fusion methods and its fault diagnosis accuracy is higher than that of the state-of-the-art methods.

The rest of the paper is summarised below. Section 2 introduces the related theories. Section 3 describes the proposed method. Section 4 conducts experimental validation and analysis. Section 5 provides conclusions and outlook.

2. Related theories

2.1. Convolutional neural network (CNN)

CNN is widely used in the field of fault diagnosis, which consists of the convolution layer, pooling layer and fully connected layer [13,27]. The convolutional layer uses a convolutional kernel to perform the sliding convolution operation on the input data. The convolutional layer has the parameter sharing property thereby reducing the number of parameters for the model, which can be described as follows:

$$X^{l+1}\left(w_{k,i}^{l}, b_{k,i}^{l}\right) = Conv(*) = \sum_{k=0}^{k-1} w_{k,i}^{l} \cdot b_{k,i}^{l} \tag{1}$$

where l denotes the number of convolutional layers, $w_{k,i}^{l}$ denotes the weight of the l-th convolutional layer, k denotes the k-th element of the l-th convolutional layer, and i denotes the neuron of the k-th element in the l-th convolutional layer. $Conv(*)$ denotes the convolutional operation, and $b_{k,i}^{l}$ denotes the bias of the l-th convolutional layer.

2.2. Kolmogorov-Arnold network (KAN)

KAN is a deep neural network with powerful function approximation capability [28]. Because of its ability to approximate arbitrary continuous functions, it shows great advantages when dealing with complex nonlinear data. In addition, since its construction is based on rigorous mathematical theorems, it has better interpretability [29].

The superposition theorem of Kolmogorov is used to construct KAN, the structure of which is shown in Fig. 1. The superposition theorem of Kolmogorov indicates that a univariate continuous function $\varepsilon_{p,q}$ of the function $f(x_1, x_2, \ldots, x_v)$ is required to satisfy the following equation:

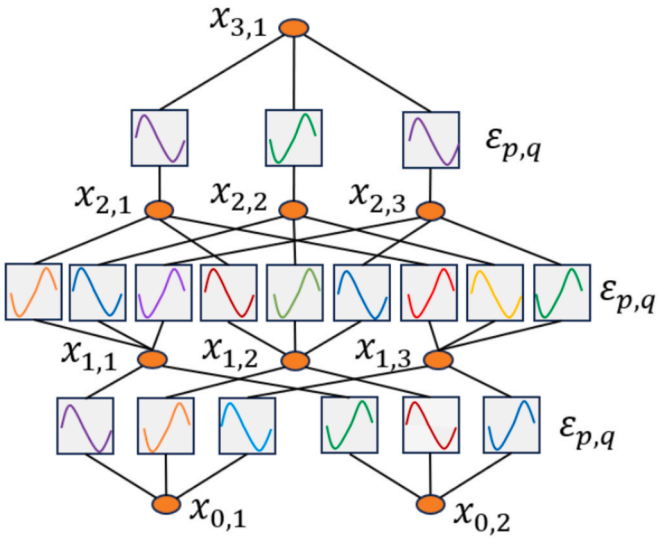


Fig. 1. Structure of KAN.

$$f(x_1, x_2, ..., x_v) = \sum_{p=0}^{2v+1} \varepsilon_{p,q} \left(\sum_{q=1}^{v} \varepsilon_{p,q}(x_p) \right) \tag{2}$$

where v is the number of independent variables for the multivariate continuous function $f(\cdot)$. p and q denote the nodes of the KAN. $p \in \{0, 1, 2, ..., 2v+1\}$, $q \in \{1, 2, 3, ..., n\}$.

2.3. Transformer

Transformer consists of an encoder and a decoder and they have the same structure: multi-head self-attention mechanism (MSA), feed-forward neural network, residual connection, and normalization [30]. MSA captures global dependencies between different positions for the input sequence by many different attention heads. In addition, MSA can reduce the training time of the model through parallel computation. The formula of MSA is described as follows:

$$MSA(Q, K, V) = Concat(head_1, head_2, ..., head_s) \tag{3}$$

$$head_i = Att(QW_q, KW_k, VW_v) \tag{4}$$

where Q is the query matrix, K is the key matrix, V is the value matrix, $Concat(\cdot)$ is the fusion operation. W_q is the query matrix weight, W_k is the key matrix weight, and W_v is the value matrix weight. $Att(\cdot)$ is the self-attention operation. $head_i$ is the i-th self-attention head.

3. The proposed fault diagnosis framework

3.1. Multi-sensor data fusion strategy

The signals of sensors at different locations can represent the overall operating conditions of mechanical equipment for rotating machinery. Therefore, it is of great significance to conduct multi-sensor fault diagnosis research on rotating machinery. However, sensor signals from different locations may introduce redundant information. How to retain important fault information and eliminate redundant information is key in the field of rotating machinery fault diagnosis. In this paper, a data level fusion strategy is designed by PCA to generate the dataset for multi-sensor data fusion. Specifically, the sensor data from different locations are selected. For example, the current signal of the motor, the vibration signal of the test equipment, and the torque signal of the drive shaft. Then, PCA is used to downscale the sensor data at different locations and generate three main components. Finally, CWT is applied to the three main components to generate R-channel images, G-channel images and B-channel images, and they are fused to generate RGB images. The specific process of data processing is described in the experimental section.

3.2. Fourier convolution (FC)

The convolution operation is the linear operation in CNN, which samples the input data by convolution kernel and sums the obtained elements. The process of convolution operation is shown in Fig. 2(a). As can be seen from the figure, the convolution operator obtained from the convolution operation is linear, which affects the nonlinear feature extraction capability of CNN. Therefore, the activation function is used to improve the nonlinear feature learning ability for the convolutional layer. However, in some complex

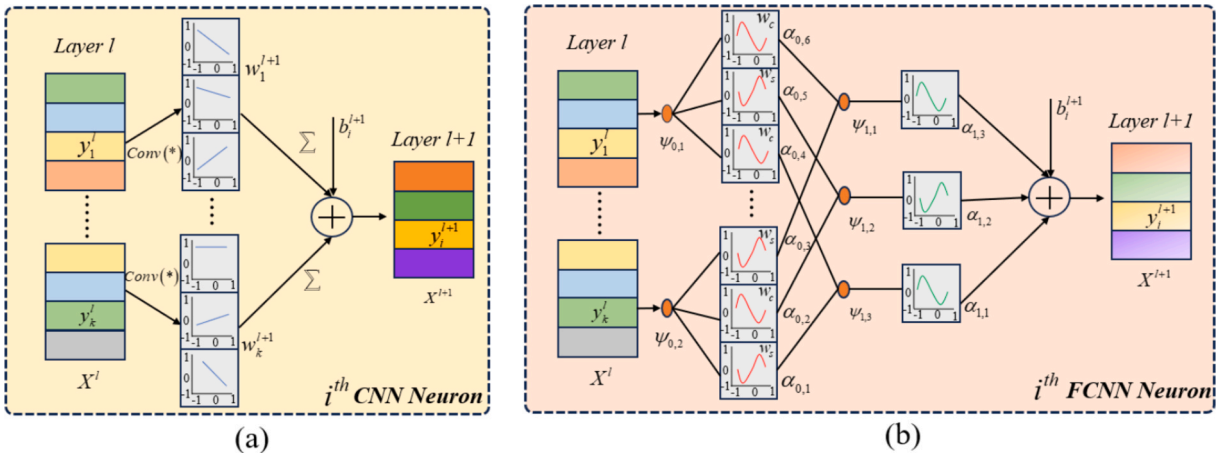


Fig. 2. Schematic structure of convolution. (a) Traditional convolution. (b) Fourier convolution.

fault diagnosis tasks, improving nonlinearity is insufficient by introducing activation functions. Inspired by Kolmogorov-Arnold Network (KAN), KAN is integrated into CNN to improve nonlinear feature extraction. In addition, since the input data of the model contains time–frequency features, the one-dimensional functions in KAN are replaced by sine and cosine functions to obtain more Fourier features. The designed FC structure is shown in Fig. 2(b).

Firstly, the one-dimensional linear base functions in the KAN are replaced with the sine and cosine in Fig. 2(b), which can be described as follows:

$$x = w_c \cdot \cos(x) + w_s \cdot \sin(x) \tag{5}$$

where w_c and w_s denote the initialised weights of the sine and cosine for transforming the input features. $\cos(\cdot)$ and $\sin(\cdot)$ denote the cosine and sine functions, respectively.

Secondly, the neuron activation value can be expressed as the sum of all the activation values from the previous layers, and the process can be described as follows:

$$\psi_{l+1} = \sum_{i=1}^{n} \alpha_{l,j}\left(w_c \cdot \cos(x_j) + w_s \cdot \sin(x_j)\right) \tag{6}$$

where l denotes the number of layers in which the neurons are located, j denotes the $j-$th base function, and n denotes the total number of neurons, and $\alpha_{l,j}(\cdot)$ is the basis function for Fourier convolution.

Finally, the weights of FC can be expressed as follows:

$$w_{FConv} = \sum_{z}^{2z+1} \psi_z \left(\sum_{j=1}^{n} a_z\left(w_c \cdot \cos(x_j) + w_s \cdot \sin(x_j)\right) \right) \tag{7}$$

where z denotes the total number of layers.

The bias term is introduced after obtaining the FC weights. Thus, the FC can be described as follows:

$$FConv\left(w_{k,i}^{l}, b_{k,i}^{l}\right) = \sum_{k=0}^{k-1} w_{k,i}^{l} \cdot b_{k,i}^{l} \tag{8}$$

where $w_{k,i}^{l}$ denotes the weight of the k-th element in layer l, and $b_{k,i}^{l}$ denotes the bias of the k-th element in layer l.

FC not only enhances the nonlinear capability of the model, but also captures local detail features for the input data. In addition, the introducing of sine and cosine functions in KAN can effectively mine the periodic and frequency fault features among the input features.

3.3. *Variational auto-encoding transformer (VAETransformer)*

Transformer is capable of global modelling for input features to obtain global fault features. However, Transformer is prone to overfitting under small sample conditions, which leads to the decrease of the diagnostic performance for the model. In order to make the model satisfy the different fault diagnosis tasks, VAETransformer is designed, and its structure is shown in Fig. 3. The left section of Fig. 3 depicts the input and token processing module. The middle section shows the Transformer layer, which consists of normalization, variational autoencoder attention, and multilayer perceptron. The right section provides the detailed description of the

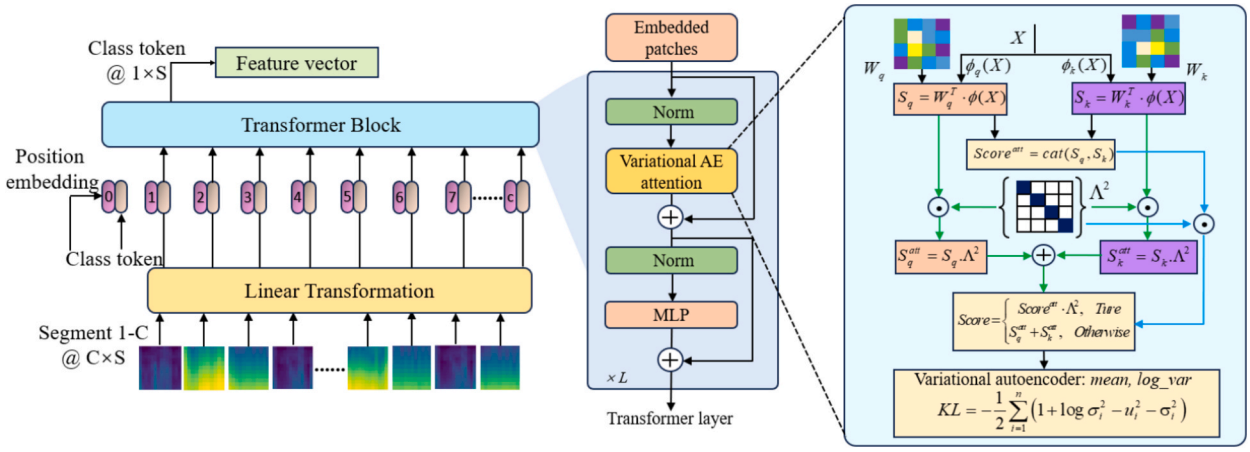


Fig. 3. Structure of VAETransformer.

variational autoencoder attention module. VAETransformer not only can provide global feature modeling but also enhances the robustness and generalization of feature learning.

The main difference between VAETransformer and Transformer is the attention mechanism. Transformer uses the multi-head self-attention mechanism for feature learning at different levels. However, the sparsity of Transformer is further amplified under small sample conditions, which limits the ability to capture feature relationships at different levels. In addition, some of the attention heads cannot learn critical information, leading to the inefficient use of computational resources. Therefore, the variational auto-encoding attention mechanism (VAE) is designed to replace the MSA, which consists of the following five steps.

Step 1: The query and key matrices are $\phi_q(X)$ and $\phi_k(X)$. The low-rank matrix is used to define the weight matrices W_q and W_k for the query and key matrices, respectively. Computing attention scores often involves matrix multiplication, which carries high computational complexity. Low-rank matrices not only enable dimensionality reduction of input features and reduce computational

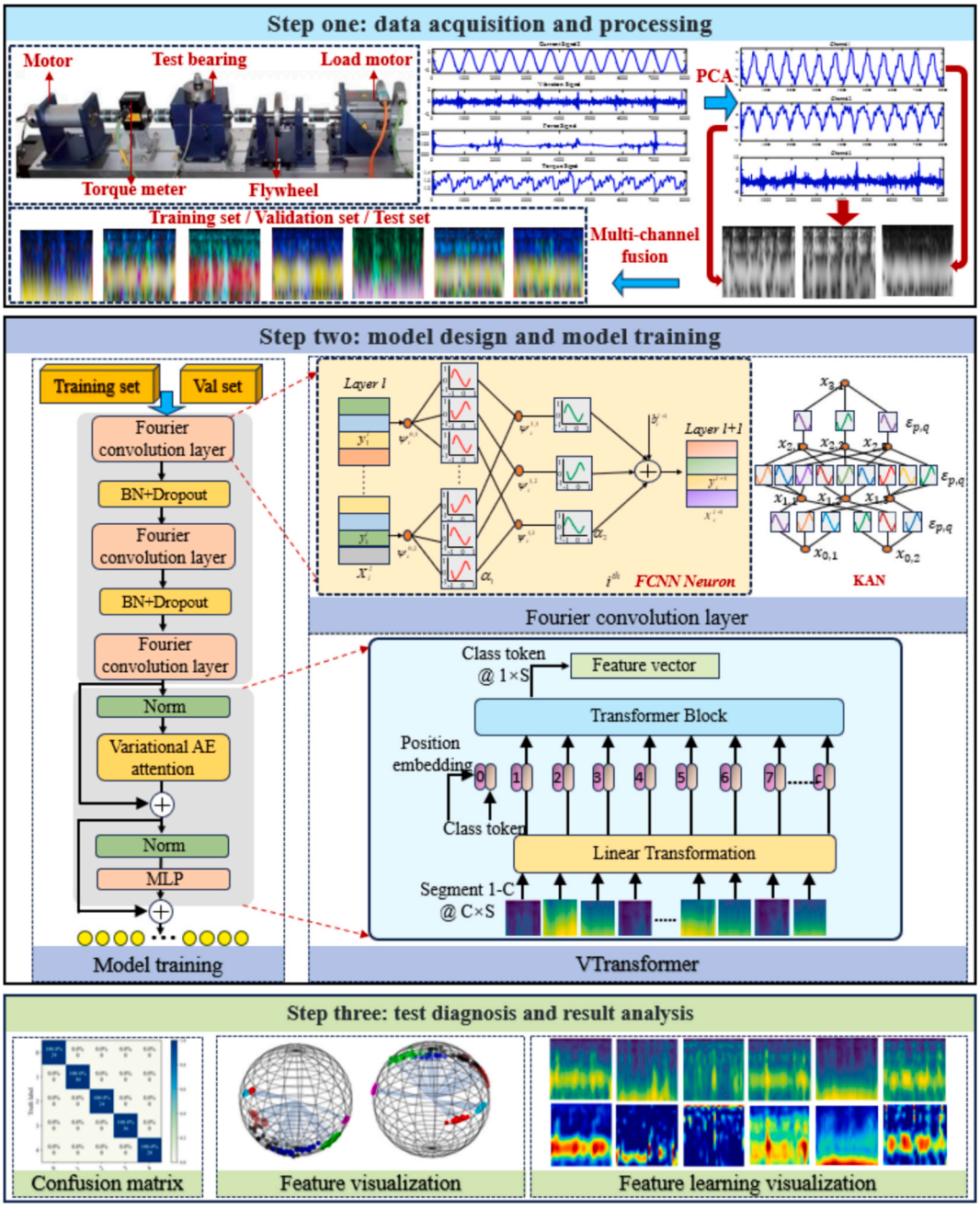


Fig. 4. The fault diagnosis framework of FCTransformer.

load, but also preserve the critical information of input features [31]. The mathematical formulae for W_q and W_k are described as follows:

$$\begin{cases} W_q = U_q \cdot B_q^T \\ W_k = U_k \cdot B_k^T \end{cases} \tag{9}$$

where U_q and U_k are the basis matrices. B_q and B_k are the coefficient matrices, $U_q, U_k \in R^{d\times r}$, $B_q, B_k \in R^{r\times r}$, and d are the lengths of the input features, and r is the dimension of the low-rank matrix.

Step 2: Attention score is obtained by the weight matrices of the query and key matrices, which can be described as follows:

$$\begin{cases} S_q = W_q^T \cdot \phi_q(X) \\ S_k = W_k^T \cdot \phi_k(X) \end{cases} \tag{10}$$

where W_q^T and W_k^T denote the transpose matrices of the query matrix and key matrix, respectively.

Step 3: The diagonal matrix Λ is defined, which scales the attention scores of the query matrix and the key matrix to control the attention weight distribution.

Step 4: There are two operations for the attention scores of the query matrix and the key matrix. One is to fuse the attention scores and then perform dot product with the diagonal matrix. The other is to perform dot product with the diagonal matrices and then perform the summing operation. The process can be described as follows:

$$Score^{att} = cat(S_q, S_k) \tag{11}$$

$$S_q^{att} = S_q \cdot \Lambda^2, \quad S_k^{att} = S_k \cdot \Lambda^2 \tag{12}$$

$$Score = \begin{cases} Score^{att} \cdot \Lambda^2, & Ture \\ S_q^{att} + S_k^{att}, & Otherwise \end{cases} \tag{13}$$

where $cat(\cdot)$ denotes the summation operation. Λ^2 denotes the square of a diagonal matrix.

Step 5: The encoder in the VAE is used to downscale the attention scores, thus reducing the redundancy of information. The attention score is reconstructed using the decoder to enhance the generalization of the model. In addition, the output features of the encoder are constrained by calculating the KL scatter to prevent the model overfitting under small sample conditions. The process can be described as follows:

$$\mu = Linear_{w_\mu, b_\mu}(Score) = Score \cdot w_\mu + b_\mu \tag{14}$$

$$\log\sigma^2 = Linear_{w_{\log}, b_{\log}}(Score) = Score \cdot w_{\log} + b_{\log} \tag{15}$$

$$KL\left(N(\mu, \sigma^2) \| N(0,1)\right) = -\frac{1}{2}\sum_{h=1}^{D}\left(1 + \log\sigma_h^2 - \mu_h^2 - \sigma_h^2\right) \tag{16}$$

where $Linear_{w_\mu, b_\mu}$ and $Linear_{w_{\log}, b_{\log}}$ are the weight and bias for mean and log variance, respectively, which are obtained from the linear layer of the VAE. μ is the mean, $\log\sigma^2$ is the log variance, and D is the number of dimension spaces.

From the above steps, it is obvious that the variational auto-encoding attention mechanism can learn rich feature information from the query and key matrices and generate the attention output through the low-rank matrix and VAE. Therefore, there is no value matrix in the variational auto-encoding attention mechanism. In addition, using VAE to regularise the model can prevent overfitting and improve the generalisation ability.

3.4. Fault diagnosis framework of FCTransformer

The proposed fault diagnosis framework of FCTransformer is shown in Fig. 4, which can include three steps.

Step 1: Data acquisition and processing. Sensor signals from different locations are collected. The different sensor data are downscaled by PCA, and CWT is used to generate R-channel, B-channel and G-channel images. The different channel images are fused to generate RGB image.

Step 2: Model design and model training. FCTransformer fault diagnosis model is constructed using FC and VAETransformer. Specifically, local periodicity and frequency fault features are first extracted by three Fourier convolution layers. Then, global dependency modelling and fine-grained feature extraction are performed using VAETransformer. Finally, the extracted high-dimensional features are mapped to the classification dimension by average pooling.

Step 3: Test diagnosis and result analysis. The test set is fed into the trained completed model to obtain fault diagnosis results. Visualisation techniques are used to analyse the results from multiple perspectives.

4. Experimental validation and analysis

In this section, the validation of FCTransformer was carried out based on two test rigs: bearing test rig and gearbox test rig. The experiment was validated on the Windows 10 operating system. The computer was equipped with a GeForce RTX 4060 video card and a 13th – Gen Intel® Core i5 – 13400F processor with a base clock speed of 2.50 GHz. The experimental environment included Python with PyTorch version 2.2.2 and MATLAB version R2020a. The CUDA version was 11.3. The code was available at https://github.com/yazhouz584-star/Fusion2.

4.1. Dataset description

Case 1: Paderborn Bearing Dataset (PBD)

The PBD dataset was provided by the University of Paderborn and the test platform was shown in Fig. 5 [32]. The platform consists of a drive motor, a measuring shaft, a flywheel, and a load motor. And five sensors were used to collect the current signal, force signal, torque signal, and vibration signal. Among them, the sampling frequency was 64 KHz. One healthy state, three outer-ring faults and three inner-ring faults were selected for multi-sensor fault diagnosis research in this experiment. The bearing fault damage modes were electric discharge and electron etching. Table 1 showed the specific fault classes and dataset division.

Case 2: Southeast University Dataset

The Southeast University (SEU) Gearbox Dataset was derived from the Southeast University Power Drive Simulation test rig [33]. The test rig consisted of a motor, a planetary gearbox, a reduction gearbox, and a brake. The multi-sensor data were collected by the acceleration sensors installed at the gearbox and motor ends. The sampling frequency was 5120 Hz. The fault classes selected for this experiment were from the gearbox including normal gears (Label 0), broken gears (Label 1), tooth wear (Label 2), cracked teeth (Label 3), and missing teeth (Label 4).

4.2. Data processing

Fig. 6 showed the signals from different sensors in the PBD dataset and the visualisation results after PCA downscaling. Specifically, Fig. 6(a) showed the signal figure for an inner ring bearing failure, Fig. 6(c) showed the signal figure for a bearing in normal condition, Fig. 6(e) showed the signal figure for an outer ring bearing failure. The differences for various bearing fault types were primarily reflected in the vibration signals and torque signals, the pulse amplitude of the vibration signal was used to indicate the extent of bearing damage. The current signals showed the current variation of the motor, which exhibited periodic characteristics. The vibration signals contained the periodic fault signals and noise information. The force signals showed the force condition of the system, which contained low frequency information and sudden change points. The torque signals showed periodic disturbances to represent load variations. Fig. 6(b), (d) and (f) showed the main components of PCA after downscaling. From the figure, it could be observed that channel 1 retained the periodicity feature, which indicated that the current signals had greater contribution to represent the state of the system. Channel 2 exhibited nonlinear trends at low frequencies, which indicated that the force and torque signals also represented the system state. Channel 3 showed significant high-frequency perturbations, which indicated that it mainly contains vibration signals.

Fig. 7 showed the greyscale and RGB fusion images of the R, G and B channels for the PBD dataset. From the figure, the R channel had smooth luminance changes, which indicated that this channel was the low-frequency component containing smoothly changing features, the G channel contained more obvious ripple-like or periodic features, and the B channel was the high-frequency component containing transient shocks, impulsive signals, and noise interferences. The RGB image was generated by fusing the greyscale images of the R, G and B channels to enhance the representation of time–frequency information.

Fig. 8 showed the signals of different sensors and the visualisation results after PCA downscaling for the SEU dataset. Specifically, Fig. 8(a) showed the signal figure for tooth wear, Fig. 8(c) presented the signal figure for missing teeth, as shown in Fig. 8(a) and Fig. 8 (c), the amplitude variations of the motor vibration signals and motor torque signals remained relatively stable. The vibration signals of the x, y and z axes for the planetary gearbox and reduction gearbox contained the periodic fault features and noise information. Fig. 8(b) and (d) showed the main components after PCA downscaling. As could be seen from the figures, the main components after PCA downscaling contained the vibration signals of the planetary gearbox and the reduction gearbox, which indicated that the PCA downscaling method could effectively retain the high-frequency information of the vibration signals.

Fig. 9 showed the greyscale and RGB fusion images of the R, G and B channels for the SEU dataset. From the figure, it could be observed that the R channel was less bright and the texture details were more blurred, which indicated that there was some noise

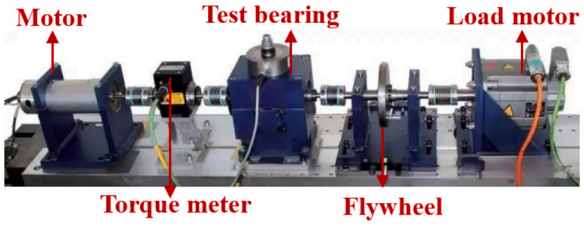


Fig. 5. The test rig for the PBD dataset.

Table 1
PBD dataset division.

Condition	Fault classes	Bearing No.	Fault damage type	Label
1500 rpm/1000 N	Healthy	K001	/	0
900 rpm/1000 N	Outer ring	KA01	Electric discharge	1
1500 rpm/1000 N	Outer ring	KA05	Electron etching	2
1500 rpm/400 N	Outer ring	KA07	Drilled hole	3
	Inner ring	KI01	Electric discharge	4
	Inner ring	KI05	Electron etching	5
	Inner ring	KI07	Drilled hole	6

interference. The texture clarification of the G channel was enhanced, which indicated that the G channel contains more fault information. The details of the B channel were further enhanced and the stripes were clearer, which indicated that this channel could capture the high-frequency information. In addition, the RGB image showed obvious coloured stripes, and the respective textures were still maintained in the fused image, which indicated that the features of each channel still maintained the certain degree of independence after fusion, and more fault information could be obtained by generating the RGB image.

4.3. Experimental setup and model parameters

The hyperparameters were set as follows: the batch size was 32, and the number of iterations was 100. The model parameters were updated by the Adam algorithm, and the learning rate was set to 0.0003. The random seed was 42. In addition, this paper established the normal sample and small sample conditions, and the total number of samples was set to 50, 100, 150, and 200, respectively. The normal samples were divided into training set, validation set, and test set according to the ratio of 0.6:0.2:0.2. The small sample conditions were divided into training set, validation set and test set according to the ratio of 0.2:0.2:0.6. The experimental results were averaged over 10 runs. Accuracy and recall were used to evaluate the fault diagnosis performance. The model parameters of FCTransformer were shown in Table 2.

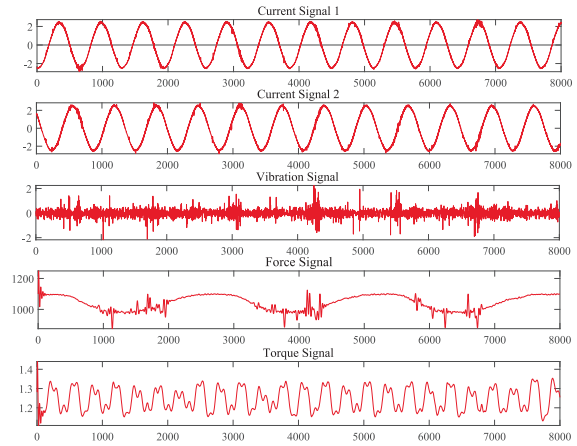
4.4. Comparison methods

In order to investigate the overall performance of FCTransformer for fault diagnosis under normal and small sample conditions, this paper compared it with nine state-of-the-art deep learning models. The details of each model were described as follows:

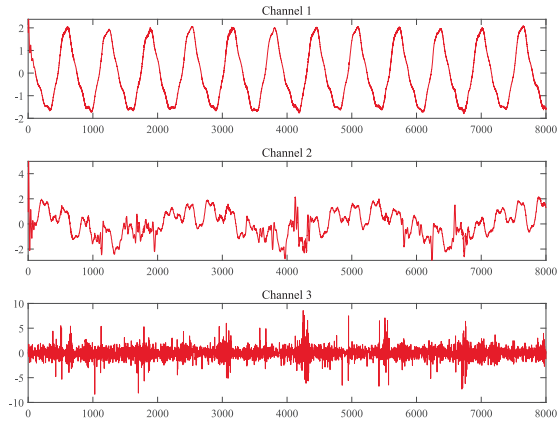
(1) IF-EDAAN [25]: IF-EDAAN was an unsupervised multi-source information domain adaptive method, which converted multi-source information into fused images by PCA, and then used CNN for feature extraction.
(2) MSF-CNN [34]: MSF-CNN was a convolutional neural network intelligent diagnosis method based on multi-sensor fusion. This method generated RGB stripe image from multi-sensor data by PCA.
(3) MSAWS [35]: MSAWS was a semi-supervised fault diagnosis method for multi-sensor data fusion, which established a weighted multi-sensor fusion strategy by the self-attention mechanism.
(4) MSCNN [36]: MSCNN was a multi-scale convolutional neural network for small sample fault diagnosis, which used Markov transfer fields (MTF) to process the input signals.
(5) PE-DCM-ViT [37]: PE-DCM-ViT was a vision transformer fault diagnosis method, which converted the input vibration signal into time–frequency image by CWT.
(6) TSSRL [38]: TSSRL was a self-supervised learning method based on Transformer, which captured global features to achieve fault diagnosis under limited labelled data.
(7) WD-KANTF [39]: WD-KANTF was a fault diagnosis method based on wavelet denoising and KANTransformer, which introduced learnable activation functions in the linear layer of the Transformer to enhance the model's nonlinear feature extraction capability.
(8) LiConvFormer [40]: LiConvFormer was a lightweight fault diagnosis method that utilised the broadcast self-attention mechanism to construct the lightweight global feature extraction module.
(9) TLMW-former [41]: TLMW-former was a trustworthy lightweight multi-expert wavelet Transformer method that designed a wavelet linear self-attention module to extract global feature information.

Besides comparing with the above state-of-the-art methods, this paper constructed different input feature images and ablation experiments, which were performed under the same benchmark as the proposed method.

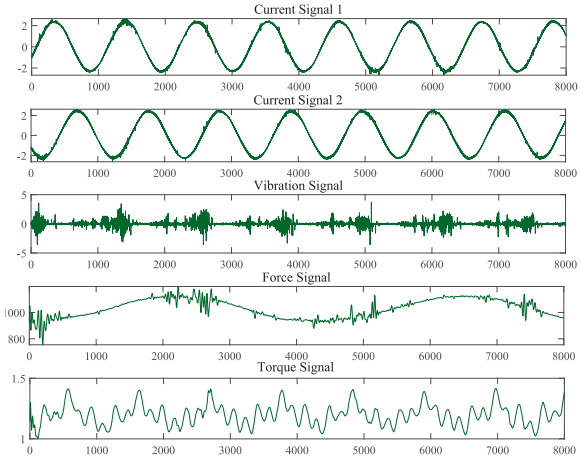
(10) STFT-CNN: Short Time Fourier Transform (STFT) was used to process the vibration signals, and standard convolution was used to replace the KAN convolution layer.
(11) STFT-CNNT: STFT was used to process the vibration signals, and standard convolution and transformer was used to perform feature extraction.
(12) STFT-FCT: STFT was used to process the vibration signals, and FCTransformer was used for feature extraction.
(13) 1D-FCTransformer: 1D-FCTransformer was the one-dimensional model of the proposed method.



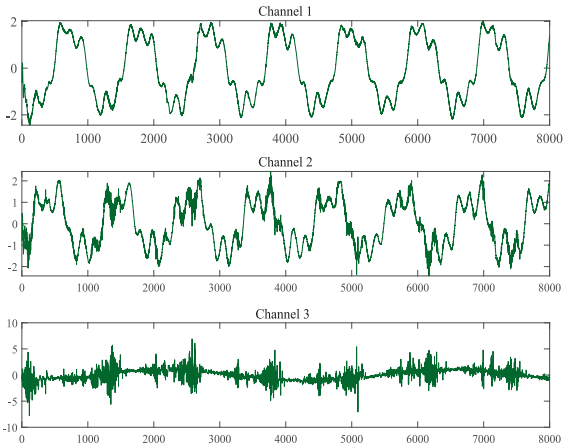
(a) KI07



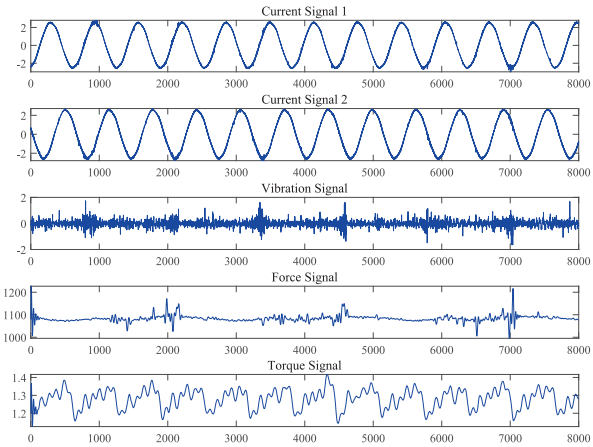
(b) KI07 principal components



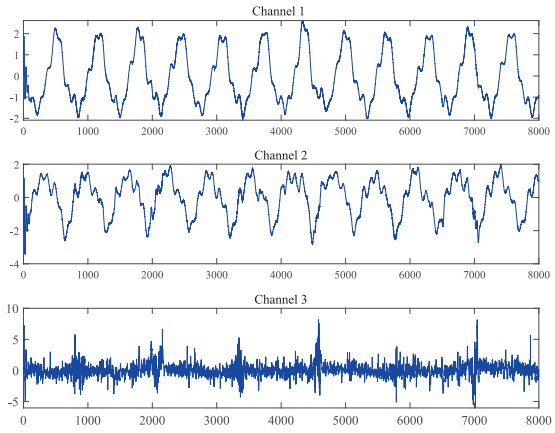
(c) K001



(d) K001 principal components



(e) KA07



(f) KA07 principal components

Fig. 6. Visualisation results of different sensor signals for the PBD dataset.

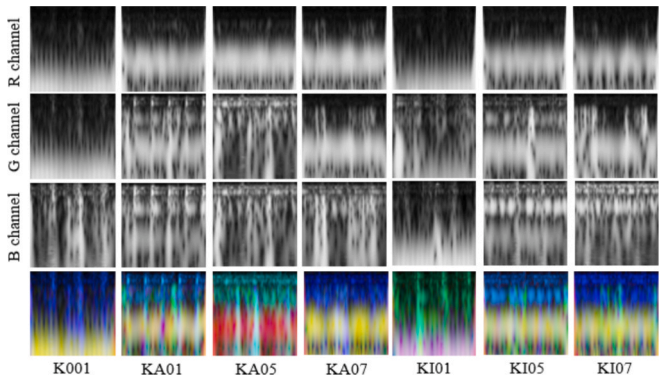


Fig. 7. RGB image of data fusion for PBD dataset.

(14) MTF-CNNT and MTF- FCT had the same structure as STFT-CNN, STFT-CNNT and STFT- FCT. However, the input feature images were Markov transfer fields.

Fig. 10 showed the different input images for the above comparison methods and ablation experiments.

4.5. Fault diagnosis results for PBD dataset

4.5.1. Fault diagnosis results under normal sample conditions

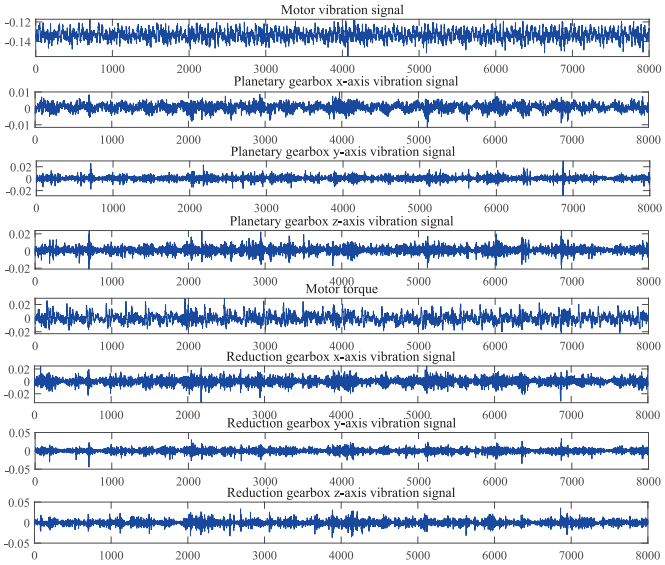
To verify the diagnosis performance of FCTransformer, sample sizes of 50, 100, 150, and 200 were selected for experiments, and the results were shown in Table 3.

In Table 3, FCTransformer performed the best in terms of accuracy, recall, Macro-F1, and F1. Specifically, FCTransformer had the average diagnostic accuracy of 98.25% at four different sample sizes, which was significantly higher than the other compared methods. Compared with IF-EDAAN and MSF-CNN, the average diagnostic accuracies of FCTransformer were improved by 4.28% and 10.64%, respectively. This indicated that CWT could obtain more time–frequency information at the data level. The average diagnostic accuracy of FCTransformer was improved by 6.51% compared to MSAWS, which indicated that more fault features could be obtained from data level fusion than feature level fusion. Compared with MSCNN and PE-DCM-ViT, the average diagnostic accuracy of FCTransformer was improved by 6.11% and 3.78%, respectively, which indicated that the diagnostic accuracy of the model could be improved by fusing the signals of sensors at different locations. In the ablation experiments, the diagnostic accuracy of FCTransformer still had high diagnostic accuracy compared to the STFT and MTF. Specifically, the average diagnostic accuracy of CNNT was 8.49% higher than the average diagnostic accuracy of CNN, which indicated that the global fault features were mined by transformer. The average diagnostic accuracy of FCTransformer was 15.87% and 7.38% higher than that of CNN and CNNT, which showed that the Fourier convolution could focus on more fault features and improve the fault diagnosis accuracy. In addition, to ensure the fair comparison, we evaluated the proposed method with one-dimensional inputs against the comparative methods with one-dimensional inputs, and the results were shown in Table 3. Compared with TSSRL, the average diagnostic accuracy of 1D-FCTransformer improved by 7.92%. Compared to WD-KANTF and TLMW-former, the average diagnostic accuracy of 1D-FCTransformer improved by 1.63% and 2.70%, respectively. This demonstrated that introducing the Fourier operator effectively improved nonlinear feature extraction capability, thereby enhancing fault diagnosis accuracy. Compared to LiConvFormer, the 1D-FCTransformer improved the average diagnostic accuracy by 9.46%. This indicated that VAETransformer could extract critical fault features to enhance fault diagnosis accuracy.

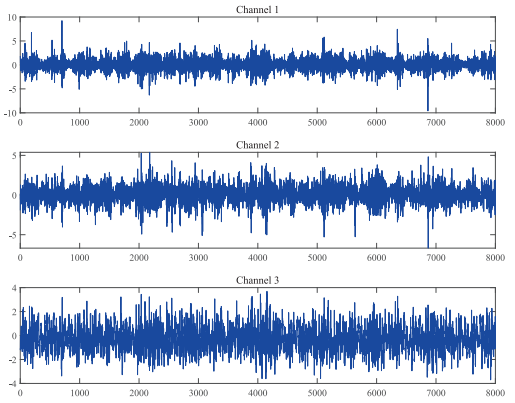
In order to observe the feature extraction ability of different methods, 150 and 200 samples were chosen to perform 3D visualisation for output features, as shown in Fig. 11. Different colours in the figure represented different classes (Label 0–6), and the clustering on the sphere intuitively reflected the classification ability of the model. As could be seen in Fig. 11, compared with the nine advanced fault diagnosis methods (a1-i1 and a2-i2), FCTransformer (p1 and p2) significantly improved the inter-class separability and intra-class compactness, which indicated that FCTransformer had more advantages in maintaining the intra-class consistency and enhancing the inter-class separability. The Transformer methods (f1-i1 and f2-i2) exhibited the scenarios where different classes were mixed, and the inter-class distances within the same classes were larger. In addition, there was significant mixing of classes for j1-n1 and j2-n2, which showed that these methods had weaker feature extraction capabilities. This further demonstrated that Fourier Convolution and VAETransformer could mine more fault features and enhance the fault identification capability.

4.5.2. Fault diagnosis results under small sample conditions

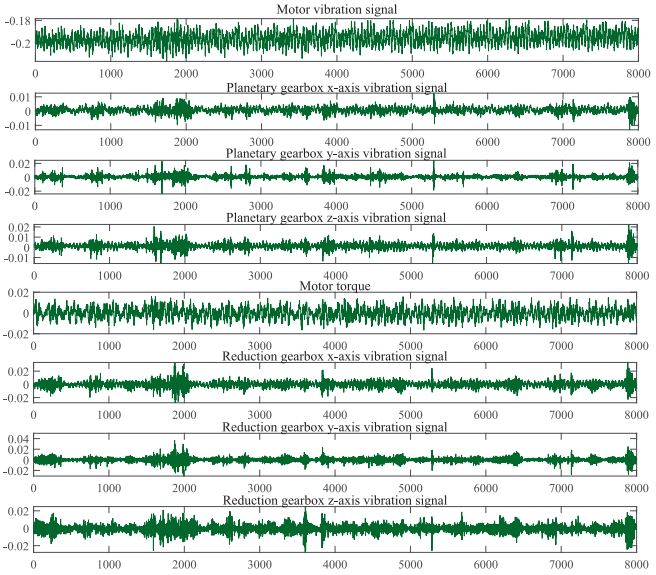
In order to further verify the effectiveness and superiority of FCTransformer under small sample conditions, sample sizes of 50, 100, 150 and 200 were selected for small sample experiments. The experimental results were shown in Table 4. In Table 4, FCTransformer performed the best in terms of accuracy, recall, Macro-F1, and F1, its average accuracy was 90.80%. This indicated that fusing sensor signals from different locations could obtain more fault information at the data level and help the model to obtain better fault diagnosis



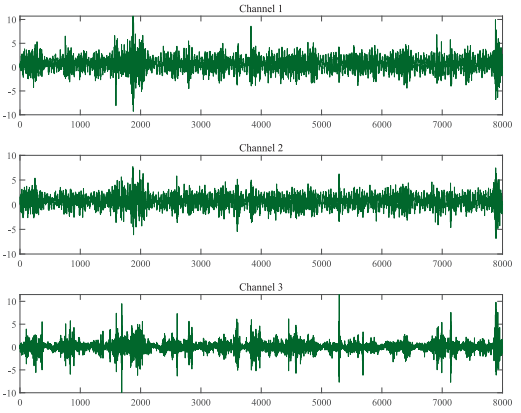
(a) tooth wear



(b) tooth wear principal components



(c) missing teeth



(d) missing teeth principal components

Fig. 8. Visualisation results of different sensor signals for the SEU dataset.

accuracy under small sample conditions. Compared with nine advanced fault diagnosis methods, the average diagnostic accuracy of FCTransformer was improved by 17.69%, 14.47%, 15.62%, 20.39%, 13.99%, 40.33%, 15.04%, 31.10%, and 21.65%, respectively. This was due to the small number of samples, which led to the lack of feature learning ability for these models and decreases in diagnostic accuracy. However, the designed FC and VAETransformer extracted more fault features from both local and global perspectives, leading to significant improvement in the accuracy of the FCTransformer. In addition, compared with the one-dimensional input model, 1D-FCTransformer showed higher average fault diagnosis accuracy. This indicated that the designed Fourier convolution and VAETransformer could extract critical fault features. The average diagnostic accuracy of FCTransformer was improved by 6.36% and 9.11% compared with SFFT-FCT and MTF-FCT, respectively. This indicated that CWT had good localization properties in both time and frequency domains, and could retain more sudden change signals and transient signals. In conclusion, it could be observed from Table 4 that the overfitting phenomenon occured due to the insufficient feature learning ability under the small sample condition, resulting in the decrease in the diagnostic accuracy. However, FCTransformer still had the high fault diagnosis accuracy, which was due to the designed Fourier Convolution could capture more Fourier features to enhance the feature learning ability of the model.

In order to further demonstrate the feature extraction capability of the model under small sample conditions, 150 and 200 samples

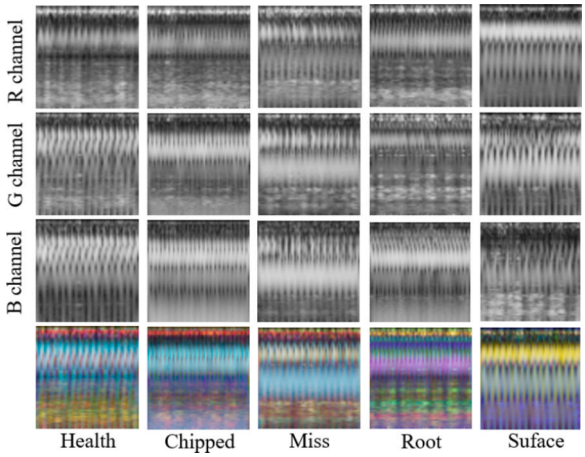


Fig. 9. RGB image of data fusion for SEU dataset.

Table 2
Parameters of FCTransformer.

Name of the layer		Output size
Fourier convolutional layer	Conv 16@ [3 × 3] BN, ReLU, Dropout (0.1)	[32, 16, 15 × 15]
Maxpooling	Size = (2, 2), Stride = (2, 2)	
Fourier convolutional layer	Conv 32@ [3 × 3] BN, ReLU, Dropout (0.1) Size = (2, 2), Stride = (2, 2)	[32, 32, 6 × 6]
Fourier convolutional layer	Conv 128@ [3 × 3] BN, ReLU, Dropout (0.1) Size = (2, 2), Stride = (2, 2)	[32, 128, 2 × 2]
VAETransformer	Number of attention heads 4 Number of embedded dimensions 128	[128, classes]

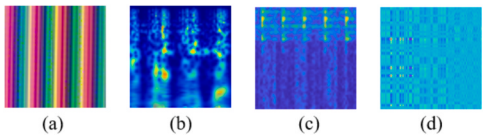


Fig. 10. The different input images (a) strip image, (b) CWT image, (c) STFT image, (d) MTF image.

were selected for 3D visualisation and the results were shown in Fig. 12. In Fig. 12, compared with nine state-of-the-art fault diagnosis methods (a1-i1 and a2-i2), FCTransformer (p1 and p2) could cluster faults of the same classes together and separate faults of different classes, which indicated that FCTransformer had great feature extraction capability, and could fully mine fault features under small sample conditions. Comparing l1, l2, n1 and n2, it could be observed that different data processing methods had different effects. Specifically, more feature information was missing when using the STFT to generate RGB images, resulting in the worst clustering effect. MTF missed feature information next, but there was overlapping of fault features under different labels. CWT could retain more time–frequency features, so it had the highest fault diagnosis accuracy and the best clustering effect under small sample conditions.

In addition, to visually demonstrate the model's identification capabilities for each fault class, we selected 30/30/90 for plotting the ROC curves, as shown in Fig. 13, the horizontal axis represented the false positive rate, while the vertical axis denoted the true positive rate. The closer the curve was to the upper-left corner, it meant the method had a lower false detection rate. In Fig. 13, the ROC curves of FCTransformer for all fault classes were significantly close to the upper-left corner, which realized the balance between high true positive rate and low false positive rate under multi-class conditions. This indicated that the proposed method had good identification accuracy for different fault classes, with both missed detection and false detection rates at low levels. The curves of IF-EDAAN, PE-DCM-ViT, STFT-FCT, MTF-FCT, and 1D-FCTransformer were next closest to the upper-left corner. These methods maintained high true positive rates for most fault classes, but false positive rates increased for some complex faults. This indicated that these methods could effectively identify most faults, but there was some risk of misclassification when encountering subtle faults. The curves for TSSRL, LiConvFormer, and TLMW-former were far from the top-left corner, and the curves for some classes were close to the diagonal line. This indicated that the classification performance of these methods was close to random levels. In small sample scenarios, the rates of

Table 3
Fault diagnosis results of PBD dataset under normal samples.

Methods	Training/validation/test							
	30/10/10		60/20/20		90/30/30		120/40/40	
	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall
IF-EDAAN	91.50%	91.18%	92.69%	92.01%	93.94%	93.90%	97.75%	97.75%
MSF-CNN	81.03%	80.05%	87.72%	86.85	89.03%	88.05%	92.63%	92.56%
MSAWS	86.77%	86.37%	90.83%	90.67%	93.43%	93.40%	95.91%	95.67%
MSCNN	87.62%	86.28%	91.17%	91.56%	94.61%	93.13%	95.16%	95.41%
PE-DCM-ViT	91.82%	92.09%	94.78%	94.68%	95.11%	94.82%	96.18%	96.01%
TSSRL	60.00%	56.62%	90.51%	90.63%	96.53%	96.50%	96.75%	96.18%
WD-KANTF	85.13%	83.76%	91.68%	89.25%	95.71%	95.63%	96.43%	96.51%
LiConvFormer	75.71%	75.48%	79.29%	81.90%	90.48%	90.46%	92.14%	92.66%
TLMW-former	77.72%	77.27%	90.20%	90.12%	98.48%	98.43%	98.28%	98.21%
STFT-CNN	71.96%	73.53%	82.81%	82.24%	86.67%	86.59%	88.09%	88.04%
STFT-CNNT	88.81%	88.76%	90.02%	89.79%	91.48%	91.43%	93.17%	93.12%
STFT-FCT	87.91%	87.87%	89.27%	89.21%	93.75%	93.67%	95.79%	96.25%
MTF-CNNT	90.88%	90.83%	92.15%	92.18%	94.44%	94.40%	96.14%	96.09%
MTF-FCT	91.71%	91.65%	94.89%	93.41%	96.15%	96.08%	97.37%	97.30%
1D-FCTransformer	88.57%	88.30%	92.16%	92.94%	96.81%	96.14%	97.96%	97.23%
FCTransformer	94.82%	94.80%	98.76%	98.57%	99.80%	99.73%	100.0%	100.0%
	Macro-F1	F1	Macro-F1	F1	Macro-F1	F1	Macro-F1	F1
IF-EDAAN	89.75%	90.00%	91.32%	91.41%	94.01%	93.72%	96.88%	96.93%
MSF-CNN	80.72%	80.74%	86.04%	86.42%	88.76%	88.77%	91.01%	91.06%
MSAWS	85.76%	85.96%	90.23%	90.47%	93.88%	93.53%	94.73%	94.96%
MSCNN	86.57%	87.01%	90.59%	91.23%	93.04%	93.85%	94.59%	95.04%
PE-DCM-ViT	90.00%	90.12%	93.36%	93.37%	94.18%	94.30%	95.61%	95.39%
TSSRL	56.24%	59.47%	87.53%	88.84%	95.43%	93.71%	95.59%	96.09%
WD-KANTF	83.06%	83.40%	90.00%	90.50%	93.88%	93.76%	95.32%	95.44%
LiConvFormer	74.28%	74.79%	78.20%	77.87%	89.50%	89.00%	91.88%	92.07%
TLMW-former	76.07%	76.20%	89.14%	89.75%	97.58%	97.63%	95.58%	97.47%
STFT-CNN	69.50%	68.67%	80.94%	81.18%	84.25%	86.21%	86.71%	86.68%
STFT-CNNT	86.20%	85.88%	88.09%	87.47%	89.78%	90.78%	93.12%	93.17%
STFT-FCT	85.77%	86.84%	87.38%	88.62%	91.07%	91.46%	94.26%	94.99%
MTF-CNNT	88.79%	89.55%	91.25%	90.56%	91.56%	92.35%	95.13%	94.93%
MTF-FCT	90.55%	89.37%	93.44%	93.33%	93.99%	95.71%	96.50%	94.76%
1D-FCTransformer	86.85%	88.34%	92.03%	92.00%	96.07%	96.21%	97.08%	97.39%
FCTransformer	94.03%	94.19%	97.73%	97.90%	99.48%	99.60%	100.0%	100.0%

missed detections and false positives were difficult to control.

4.6. Fault diagnosis results for SEU dataset

4.6.1. Fault diagnosis results for normal sample conditions

To further validate the generalization performance of FCTransformer, experimental validation was carried out using the SEU dataset. The experimental setup was the same as in Section 4.5, as shown in Table 5.

Table 5 used accuracy, recall, Macro-F1, and F1 scores to evaluate model performance. In Table 5, the fault diagnosis accuracy of FCTransformer was the highest, and its average diagnosis accuracy was 99.58%, which indicated that the designed FC could effectively extract the periodic fault features, thus helping the model to better identify the fault classes. Compared to IF-EDAAN and MSF-CNN, the average diagnostic accuracy of FCTransformer was improved by 21.39% and 24.58%, respectively. This indicated that CWT could obtain more time–frequency features, which improved the fault diagnosis accuracy. Compared with MSAWS, MSCNN and PE-DCM-ViT, the average diagnostic accuracy of FCTransformer was improved by 12.40%, 27.99% and 14.87%, respectively. Compared with directly processing the vibration signals by CWT and MTF, generating RGB images at different locations could obtain more fault information. In addition, to ensure the fair comparison, we evaluated the proposed method with the one-dimensional inputs against the comparative methods with one-dimensional inputs, and the results were shown in Table 5. Compared with TSSRL, the average diagnostic accuracy of 1D-FCTransformer improved by 14.06%. Compared to TLMW-former, the average diagnostic accuracy of 1D-FCTransformer increased by 4.65%, which indicated that introducing the Fourier operator effectively enhanced nonlinear feature extraction capability. Compared to LiConvFormer, the average diagnostic accuracy of 1D-FCTransformer increased by 9.02%, which indicated that VAETransformer could extract key fault features to enhance fault diagnosis accuracy.

Fig. 14 showed the confusion matrix plot for 150 samples, and the difference in fault identification ability of each model could be visualised in Fig. 14. The confusion matrix presented the classification results in terms of true and predicted labels. Among them, the darker diagonal colour indicates the higher model identification accuracy. As could be seen from the figure, FCTransformer achieved 100% fault diagnosis accuracy on all classes, which indicated that FC and VAETransformer could significantly improve the model

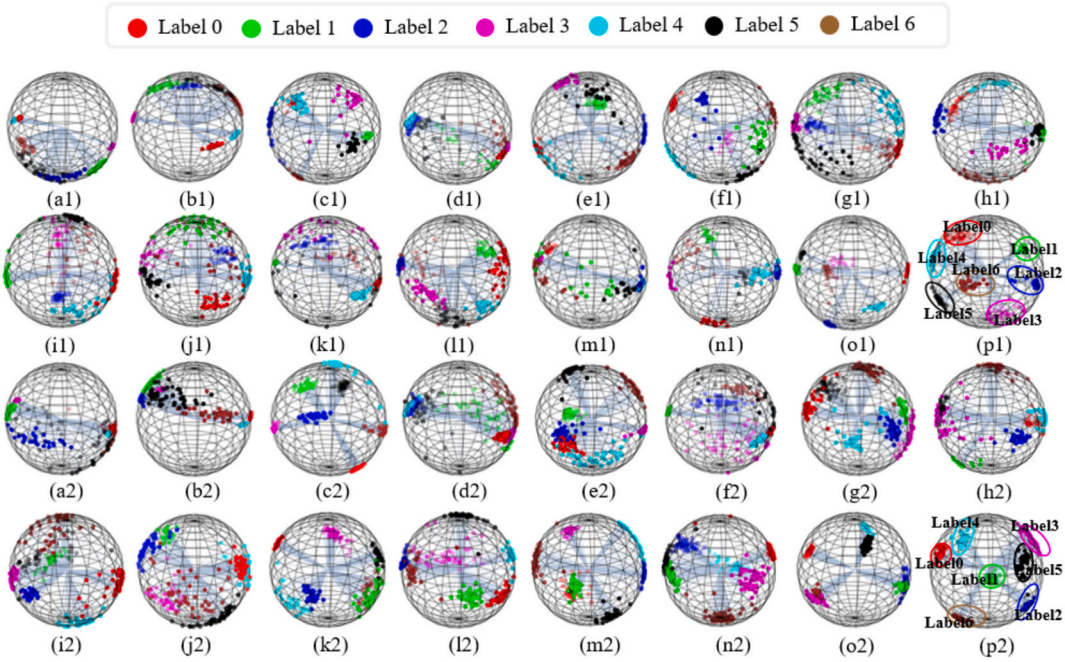


Fig. 11. Fault diagnosis visualisation results for normal sample conditions. a1-l1 are 150 samples. a2-l2 are 150 samples. (a1) IF-EDAAN, (b1) MSF-CNN, (c1) MSAWS, (d1) MSCNN, (e1) PE-DCM-ViT, (f1) TSSRL, (g1) WD-KANTF, (h1) LiConvFormer, (i1) TLMW-former, (j1) SFFT-CNN, (k1) SFFT-CNNT, (l1) STFT-FCT, (m1) MTF-CNNT, (n1) MTF-FCT, (o1) 1D-FCTransformer, (p1) FCTransformer.

feature extraction ability to obtain the best fault diagnosis accuracy. WD-KANTF and MTF-KACNT diagnostic results were second best, which indicated that different data processing methods had different effects on model diagnostic results. STFT-FCT and MTF-FCT diagnostic results were the next best, which indicated that different data processing methods had different effects on the model diagnosis results. From Fig. 14(h) and (k), it could be observed that the STFT obtained more fault information. MSF-CNN and MSAWS had high misclassification in several classes, which showed that their extraction capability was insufficient. In addition, IF-EDAAN and PE-DCM-ViT had higher fault identification accuracy, but still had the problem of lower identification for some classes.

4.6.2. Fault diagnosis results under small sample conditions

In order to further validate the fault diagnosis performance of FCTransformer, small sample experiments were conducted on the SEU dataset, and the results were shown in Table 6. In Table 6, the fault diagnosis accuracies of all methods were reduced compared to the fault diagnosis accuracies under normal sample conditions. This was due to the fact that the small number of training samples led to insufficient feature learning ability or overfitting phenomenon for the model, which reduced the fault diagnosis performance of the model. However, compared to the comparison methods, the average fault diagnosis accuracy of FCTransformer was 96.10%, which was much higher than the comparison methods. This indicated that FC could mine more periodic fault features under small sample conditions, thus improving the feature learning capability. In addition, compared with STFT-CNNT, the average fault diagnosis accuracy of STFT-FCT was improved by 11.86%, which indicated that VAETransformer could prevent the model overfitting phenomenon. In addition, compared to the lightweight LiConvFormer, 1D-FCTransformer improved the average fault diagnosis accuracy by 7.98%, which indicated that the lightweight model's simpler structure resulted in insufficient feature learning. Compared to TLMW-former, KACNT increased the average fault diagnosis accuracy by 0.24%.

In order to visualise the classification of fault classes, 150 samples were selected for confusion matrix visualization, and the results were shown in Fig. 15. In Fig. 15, CNN had obvious misclassification, but CNNT and FCTransformer performed better in fault identification. Specifically, IF-EDAAN inter-class confusion was more serious, especially label 0, label 2, and label 3 had high misclassification. This indicated that the feature learning ability of the model was insufficient under small sample conditions. There was serious confusion between label 0 and label 2 for MSF-CNN, which indicated that the learning ability of the CNN was limited. Although MSF-CNN fused sensor information from different locations, there was still misclassification in the face of high similarity classes. The identification of label 1 and label 2 was only 9.1% and 48.3% for MSAWS, which was due to the lack of global feature learning capability for the model. There were still some misidentifications for label 2 and label 3 in MSCNN, which indicated that MSCNN lacked the ability to identify combined fault classes. In PE-DCM-ViT, transformer did not sufficiently decouple the inter-class similarity features, which led to low identification for label 2 and label 3. This indicated that PE-DCM-ViT was prone to overfitting when trained under small sample conditions. In TSSRL and TLMW-former, labels 0 and 1 had high misclassification rates, which indicated that using Transformer for global dependency modelling had overlooked local fault features. In LiConvFormer, label 2 was completely misclassified, which suggested that lightweight models had limited diagnostic performance under small sample fault diagnosis conditions.

Table 4
Fault diagnosis results of PBD dataset under small sample conditions.

Methods	Training/validation/test							
	10/10/30		20/20/60		30/30/90		40/40/120	
	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall
IF-EDAAN	57.60%	57.51%	68.17%	67.13%	78.03%	77.93%	88.62%	88.13%
MSF-CNN	54.66%	54.14%	74.93%	74.88%	83.87%	82.11%	91.85%	91.69%
MSAWS	59.51%	59.37%	70.36%	70.17%	82.55%	82.29%	88.27%	88.26%
MSCNN	37.51%	37.45%	77.50%	75.48%	81.98%	81.55%	86.48%	85.47%
PE-DCM-ViT	66.26%	65.58%	74.32%	73.77%	81.30%	81.07%	85.33%	85.28%
TSSRL	36.19%	37.06%	45.48%	45.08%	53.97%	54.68%	65.35%	63.10%
WD-KANTF	64.67%	64.50%	73.61%	73.32%	77.56%	78.04%	84.29%	84.23%
LiConvFormer	42.38%	42.35%	51.67%	51.58%	60.48%	61.02%	81.36%	81.31%
TLMW-former	52.47%	53.09%	62.76%	62.46%	74.13%	73.72%	84.33%	84.05%
STFT-CNN	60.82%	59.96%	66.41%	65.81%	82.67%	82.55%	84.42%	84.33%
STFT-CNNT	64.53%	64.33%	76.01%	75.14%	84.06%	83.83%	87.13%	83.06%
STFT-FCT	72.33%	71.43%	83.42%	81.36%	90.89%	90.32%	93.17%	92.57%
MTF-CNNT	58.08%	57.76%	73.58%	72.64%	85.24%	84.92%	86.13%	86.25%
MTF-FCT	65.04%	65.02%	77.98%	77.78%	89.95%	89.44%	93.80%	93.71%
1D-FCTransformer	64.29%	64.07%	70.00%	71.49%	83.39%	83.54%	88.05%	88.37%
FCTransformer	83.09%	83.12%	87.52%	87.44%	94.82%	94.76%	97.79%	97.66%
	Macro-F1	F1	Macro-F1	F1	Macro-F1	F1	Macro-F1	F1
IF-EDAAN	56.45%	56.37%	66.91%	67.11%	76.49%	77.74%	88.77%	87.84%
MSF-CNN	53.18%	53.48%	73.08%	73.21%	82.76%	82.71%	92.72%	92.49%
MSAWS	59.17%	59.24%	69.21%	69.59%	83.38%	82.49%	87.23%	86.82%
MSCNN	37.05%	37.38%	75.98%	76.67%	80.92%	80.40%	85.01%	84.61%
PE-DCM-ViT	65.21%	65.42%	73.37%	73.38%	80.77%	80.98%	86.68%	85.72%
TSSRL	36.00%	36.09%	45.95%	46.94%	51.11%	39.66%	65.06%	66.58%
WD-KANTF	64.42%	64.75%	72.12%	72.20%	74.72%	74.87%	83.05%	83.57%
LiConvFormer	42.95%	43.08%	51.31%	51.54%	60.32%	60.97%	80.65%	80.98%
TLMW-former	53.12%	52.66%	62.06%	62.38%	74.60%	71.57%	83.55%	82.93%
STFT-CNN	59.16%	59.34%	64.87%	65.10%	81.38%	80.53%	83.08%	83.37%
STFT-CNNT	64.05%	63.93%	75.04%	75.25%	83.19%	83.36%	86.24%	87.05%
STFT-FCT	71.97%	72.83%	81.98%	82.34%	89.59%	89.32%	92.18%	93.71%
MTF-CNNT	58.00%	58.01%	72.60%	72.88%	84.28%	85.12%	86.48%	85.03%
MTF-FCT	64.22%	64.84%	77.12%	76.96%	88.02%	87.21%	93.47%	93.79%
1D-FCTransformer	63.67%	63.80%	69.49%	70.00%	83.76%	83.39%	86.34%	88.05%
FCTransformer	82.62%	82.90%	86.30%	86.39%	92.30%	92.14%	96.80%	96.20%

In addition, the time–frequency feature extraction of STFT-CNN and STFT-CNNT was insufficient, which led to the decrease of fault diagnosis accuracy. MTF-CNNT had an identification of 9.1% for label 3, which indicated that there was the risk of overfitting. The fault diagnosis accuracy of FCTransformer was 96.94%, which indicated that FC and VAETransformer had significantly improve the fault class identification ability.

In the SEU dataset, we selected the 30/30/90 for ROC curve plotting, as shown in Fig. 16. In Fig. 16, the ROC curves were significantly close to the upper-left corner for FCTransformer, this indicated that FCTransformer had excellent identification accuracy for the different fault classes, and its missed detection rate and false positive rate were at extremely low levels. In contrast, the ROC curves of IF-EDAAN, MSF-CNN, WD-KANTF, TLMW-former, STFT-FCT, MTF-FCT, and 1D-FCTransformer were farther from the upper-left corner, this was because they could effectively identify most normal faults, but they had some risks of false positives for subtle faults. The ROC curves for TSSRL, STFT-CNN, and STFT-CNNT were the farthest from the top-left corner, they were close to the diagonal line for some fault classes, this indicated that the classification performance of these methods was close to the level of random classification.

4.7. Multi-sensor data fusion analysis

The above analyses convincingly validated the effectiveness of FCTransformer, and the confusion matrix was plotted to further elaborate the effect of different sensor signals, as shown in Fig. 17. Fig. 17(a) showed the fault diagnosis results after fusing the signals from five different sensors. Fig. 17(b) showed the diagnostic results after fusion of current signals, vibration signals, force signals and torque signals. Fig. 17 (c) showed the diagnostic results after fusion of current signals, vibration signals and torque signals. Fig. 17 (d) showed the diagnostic results of the current signals. Fig. 17 (e) showed the diagnostic results of the vibration signals. Compared to the Fig. 17(b), the fault diagnosis accuracy of FCTransformer increased from 99.52% to 100.0%, this was because Fig. 17(b) lacked current signal 1, which made the fused feature image unable to fully retain all fault related information, leading to the decrease of the model's fault diagnosis accuracy. Compared to Fig. 17(c), the FCTransformer fault diagnosis accuracy was improved from 96.70% to 100.0% after fusing five sensor signals, this was because Fig. 17(c) lacked both the force signal and current signal 2. The force signal was

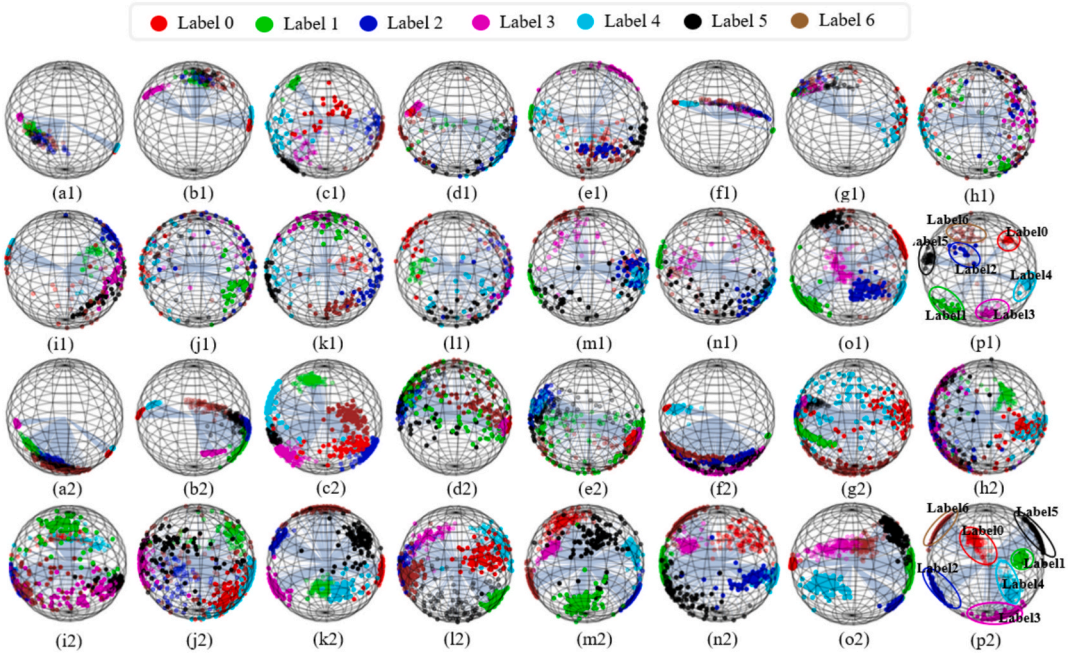


Fig. 12. Fault diagnosis visualisation results under small sample conditions. a1-l1 are 150 samples. a2-l2 are 150 samples. (a1) IF-EDAAN, (b1) MSF-CNN, (c1) MSAWS, (d1) MSCNN, (e1) PE-DCM-ViT, (f1) TSSRL, (g1) WD-KANTF, (h1) LiConvFormer, (i1) TLMW-former, (j1) SFFT-CNN, (k1) SFFT-CNNT, (l1) STFT-FCT, (m1) MTF-CNNT, (n1) MTF-FCT, (o1) 1D-FCTransformer, (p1) FCTransformer.

significant important in the entire rotating machinery monitoring system, as it directly reflected the impact of the load on the bearings. The model could not fully capture fault characteristics without the force signal, leading to reduced diagnostic accuracy.

In order to show the diagnostic results of different sensors, this paper used solid lines of different colours to connect each of the above confusion matrices. As could be seen from the red box and the red solid line in Fig. 17, the current signals provided less fault information compared to the vibration signals, resulting in the lower fault diagnosis accuracy. After fusing the vibration signals, current signals and torque signals, the fault diagnosis accuracy of the model was significantly improved. Specifically, after adding current and torque signals to the vibration signals, the diagnostic accuracies of label 0 and label 2 increased from 78.0% and 92.5% to 100.0%, respectively, which indicated that fusing different sensor signals helped to improve the fault diagnosis accuracy. After adding force signals to the vibration, current and torque signals, the diagnostic accuracies of label 4 and label 5 were improved from 88.4% and 92.2% to 100%, respectively, as shown by the orange box and orange solid line in Fig. 17. This indicated that the force signals could help the model to obtain more fault information.

Fig. 18 showed the confusion matrix results after fusion of different sensor signals for the SEU dataset. Fig. 18 (d), (e) and (f) showed the motor vibration signals, reduction gearbox x-axis vibration signals and planetary gearbox x-axis vibration signals fault diagnosis results, respectively. As could be seen from the figure, the motor vibration signals fault diagnosis results are the worst, the reduction gearbox x-axis vibration signals fault diagnosis results were the second best, and the planetary gearbox x-axis vibration signals fault diagnosis was the best. This indicated that the planetary gearbox x-axis vibration signals could represent more fault information, making it easier for the model to capture fault features. Fig. 18(c) showed the fault diagnosis results after fusion of motor vibration signals, reduction gearbox and planetary gearbox x-axis signals. As could be seen from the figure, the fusion of three differently located sensors resulted in 100% for label 2 and label 3, but there was the decrease in the fault diagnosis accuracy for label 0 and label 4, as shown by the purple solid line and the green solid line in the figure. This indicated that sensors at different locations sometimes bring redundant information, which reduced the fault diagnosis accuracy of the model. Fig. 18(b) showed the fault diagnosis results after fusion of motor vibration signals, reduction gearbox and planetary gearbox x-axis, y-axis and z-axis signals. It could be observed that the accuracy of fault diagnosis was 100% for both label 0 and 4 after fusing the vibration signals at different positions of the gearbox and the motor torque signals. Fig. 18(a) showed the fault diagnosis results after fusing eight different position sensors. From the figure, it could be observed that the fault diagnosis accuracy of the model is 100% after fusing eight different sensor signals.

4.8. Feature learning visualization

4.8.1. Gradient class activation mapping (Grad-CAM) visualization

In order to reveal the ability of Fourier convolution to capture critical feature information, Grad-CAM was used for visualization, and the results were shown in Fig. 19. Grad-CAM allowed for the intuitive observation of differences in weight gradient distributions,

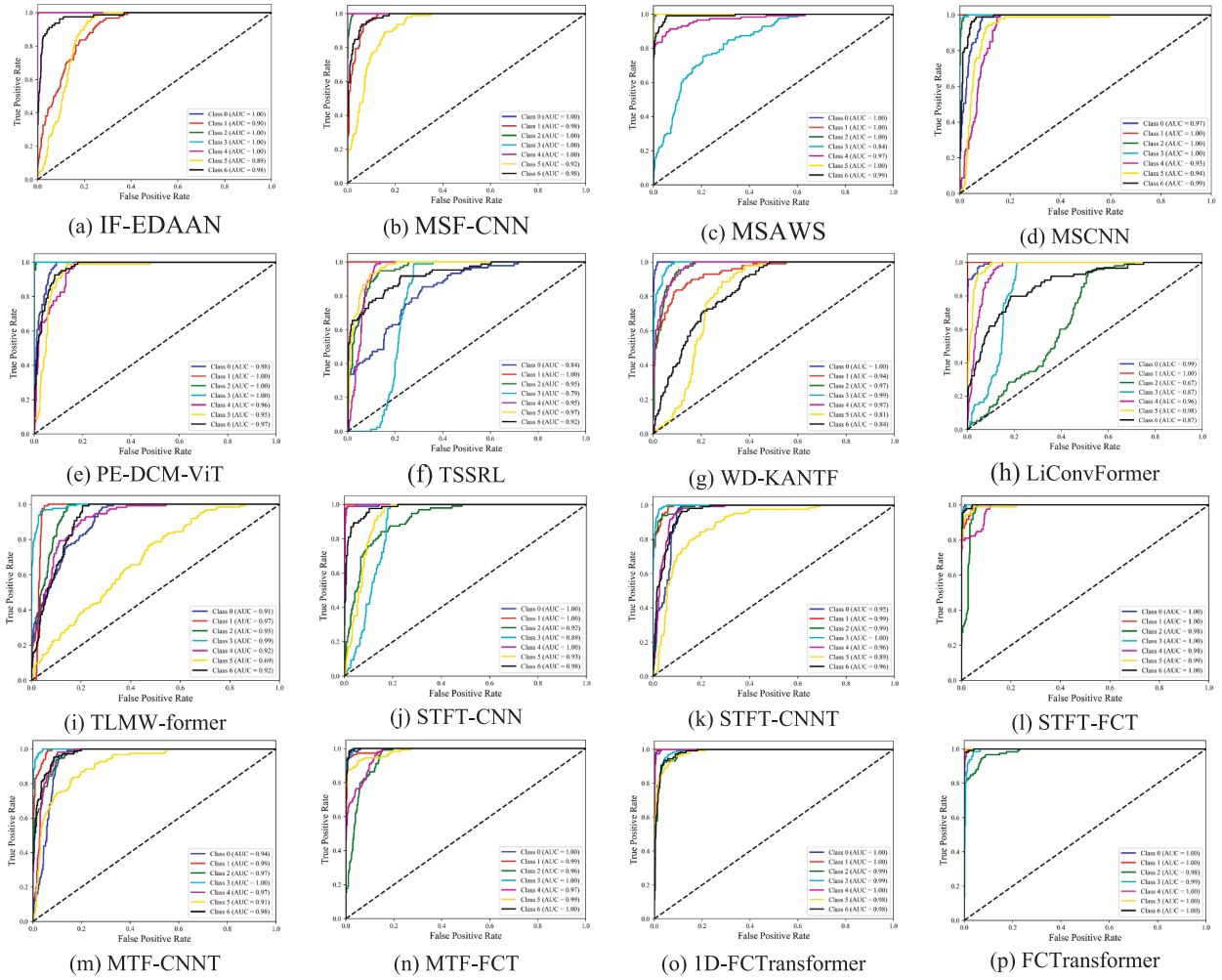


Fig. 13. The results of ROC curves for different methods on the PBD dataset.

which provided deeper insight into the internal working mechanisms of the model when processing different samples. In Fig. 19, the input image was normalised and the critical regions were marked with brighter colours (yellow). After the normalised image had gone through the first Fourier convolution layer, the important regions were marked with brighter colours (orange) in the image, which showed that FC could extract critical fault features. In addition, after the second Fourier convolution layer, the important regions were further extracted in the image, which further indicated that the designed FC autonomously extracted important features while suppressing redundant information. This performance not only helped the model to improve the fault diagnosis accuracy, but also enhanced the interpretability of the model, which was important in real-world fault diagnosis.

4.8.2. Feature learning procedure for VAETransformer

To validate the feature learning capability of VAETransformer, the PBD dataset was selected for visualization analysis. The training/validation/test splits were 30/30/90. First, we visualised the weight matrices of the query matrix and the key matrix, as shown in Fig. 20. In Fig. 20, the horizontal axis was the feature dimension index of the low-rank matrix, and the vertical axis was the position index of the token. As could be seen from the figure, there was a significant difference in the distribution of the weights for the query and key matrices, which indicated that the variational auto-encoding attention mechanism could assign different weights to the features. Specifically, the weight matrix of the query matrix assigned lower weights for most features, which indicated that the query matrix could capture local critical fault features. The values of the weight matrix for the key matrix were individually balanced and assign higher weights to most features, which indicated that the key matrix was capable of global dependency modelling.

To further validate the effectiveness of the variational autoencoder attention mechanism, the PBD dataset was again selected for visualization analysis. The training/validation/test splits were set to 30/30/90. The visualization results were shown in the Fig. 21. Heatmap colours indicated attention score values. Bright (yellow) indicated higher score values and dark (purple) indicated lower score values. In Fig. 21(a), Head0 and Head3 had balanced distribution of attention score values, which indicated that the attention mechanism performed global feature extraction. Head1 and Head2 had higher score values in the projection index1 and 3 regions,

Table 5
Fault diagnosis results of SEU dataset under normal samples.

Methods	Training/validation/test							
	30/10/10		60/20/20		90/30/30		120/40/40	
	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall
IF-EDAAN	62.70%	65.08%	72.16%	70.24%	88.32%	87.70%	89.59%	88.31%
MSF-CNN	70.01%	69.45%	72.38%	72.23%	78.46%	76.30%	79.15%	78.89%
MSAWS	79.86%	77.98%	82.16%	82.04%	92.64%	90.00%	94.05%	93.67%
MSCNN	61.51%	60.22%	65.93%	65.64%	72.50%	72.13%	86.39%	86.34%
PE-DCM-ViT	72.72%	70.61%	77.43%	77.21%	93.52%	93.44%	95.15%	95.00%
TSSRL	60.32%	60.00%	60.18%	59.13%	89.74%	89.33%	85.63%	85.00%
WD-KANTF	83.84%	82.00%	89.73%	88.03%	96.67%	96.62%	98.75%	98.33%
LiConvFormer	60.82%	58.00%	72.10%	72.00%	89.33%	89.42%	93.78%	93.19%
TLMW-former	76.93%	76.17%	80.00%	78.86%	87.09%	86.24%	89.50%	89.12%
STFT-CNN	75.62%	75.37%	76.36%	76.26%	78.66%	78.33%	89.06%	88.59%
STFT-CNNT	77.48%	77.43%	87.14%	86.41%	95.82%	95.34%	96.35%	96.37%
STFT-FCT	82.03%	81.49%	89.92%	89.57%	96.36%	96.18%	97.77%	97.65%
MTF-CNNT	70.52%	69.56%	75.83%	75.74%	82.36%	82.07%	90.00%	90.15%
MTF-FCT	75.33%	72.22%	78.06%	77.96%	85.38%	85.09%	91.11%	90.87%
1D-FCTransformer	80.00%	80.45%	86.00%	86.64%	90.64%	90.73%	95.40%	95.50%
FCTransformer	98.33%	97.14%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Macro-F1	F1	Macro-F1	F1	Macro-F1	F1	Macro-F1	F1
IF-EDAAN	61.54%	61.90%	73.66%	72.56%	86.95%	86.58%	89.73%	90.20%
MSF-CNN	71.45%	70.51%	72.65%	73.17%	76.65%	77.66%	78.05%	79.67%
MSAWS	78.57%	78.15%	81.47%	82.73%	90.66%	91.78%	92.63%	93.48%
MSCNN	58.33%	60.31%	64.44%	62.50%	71.33%	70.41%	84.56%	85.35%
PE-DCM-ViT	71.43%	73.08%	75.57%	78.19%	93.37%	93.18%	95.18%	96.94%
TSSRL	59.60%	62.67%	61.67%	63.33%	89.58%	89.93%	84.94%	85.74%
WD-KANTF	81.57%	82.00%	87.66%	87.65%	96.88%	96.08%	98.08%	97.51%
LiConvFormer	58.20%	59.56%	71.93%	71.57%	87.04%	88.89%	93.56%	93.83%
TLMW-former	75.18%	75.37%	79.58%	81.93%	86.31%	87.62%	88.51%	89.30%
STFT-CNN	74.67%	74.06%	73.33%	75.36%	79.23%	78.02%	87.62%	88.85%
STFT-CNNT	75.21%	78.19%	86.72%	87.66%	95.56%	95.48%	94.34%	96.15%
STFT-FCT	81.30%	82.35%	89.93%	90.32%	95.82%	96.19%	97.62%	96.55%
MTF-CNNT	69.74%	70.79%	75.79%	76.32%	82.42%	84.79%	90.94%	90.05%
MTF-FCT	74.15%	72.48%	78.23%	78.57%	85.05%	85.66%	90.55%	91.79%
1D-FCTransformer	80.30%	80.24%	86.40%	86.00%	90.60%	90.55%	95.44%	95.36%
FCTransformer	97.37%	98.19%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

which indicated that the attention mechanism performed capture local feature extraction. In Fig. 21(b), Head0 and Head3 performed global dependency modelling and their attention score values were relatively balanced. Head1 and Head2 assigned higher score values to projection index 8 and 1, which showed that the attentional mechanism was extracting critical fault features. In addition, the red box in the figure showed the region with the highest score value for the variational auto-coding attention mechanism. As could be seen from the figure, the region was not the same, which indicated that the variational auto-coding attention mechanism could successfully extract class discriminative features from different input samples.

4.9. Ablation experiment and parameter sensitivity analysis

To verify the impact of each module of FCTransformer in fault diagnosis performance, ablation experiments were conducted. Specifically, FCTransformer consisted of data-level multi-sensor data fusion, Fourier Convolution layer, and VAETransformer. Therefore, the ablation experiments designed were shown in Table 7 and Fig. 22.

Ablation 1: Impact of multi-sensor data fusion at the data level. After removing multi-sensor data fusion at the data level, the identification accuracy rates of the PBD dataset were 65.83%, 70.62%, 77.56%, and 85.54%, respectively. When data-level multi-sensor data fusion was introduced, the identification accuracy rates increased by 17.26%, 16.90%, 17.26%, and 12.25%, respectively. This indicated that multi-sensor data fusion could provide more fault information to enhance the model's fault diagnosis accuracy. Additionally, as shown in the table, data-level multi-sensor data fusion had the greatest impact on the model's diagnostic performance.

Ablation 2: Impact of the Fourier Convolution layer. After removing the Fourier Convolution, the identification accuracy of the PBD dataset decreased by 13.59%, 10.14%, 11.21%, and 9.15%, respectively. The identification accuracy rates of the SEU dataset decreased by 16.53%, 12.52%, 12.81%, and 7.36%, respectively. This indicated that the Fourier Convolution could extract local time–frequency features to enhance the model's fault diagnosis accuracy.

Ablation 3: Impact of VAETransformer. After removing VAETransformer, the identification accuracy of the PBD dataset decreased by 8.81%, 7.40%, 7.82% and 5.61% respectively. The identification accuracy rates of the SEU dataset decreased by 13.58%, 9.67%, 9.98%, and 4.54%, respectively. This indicated that VAETransformer could perform global dependency modelling and fine-grained

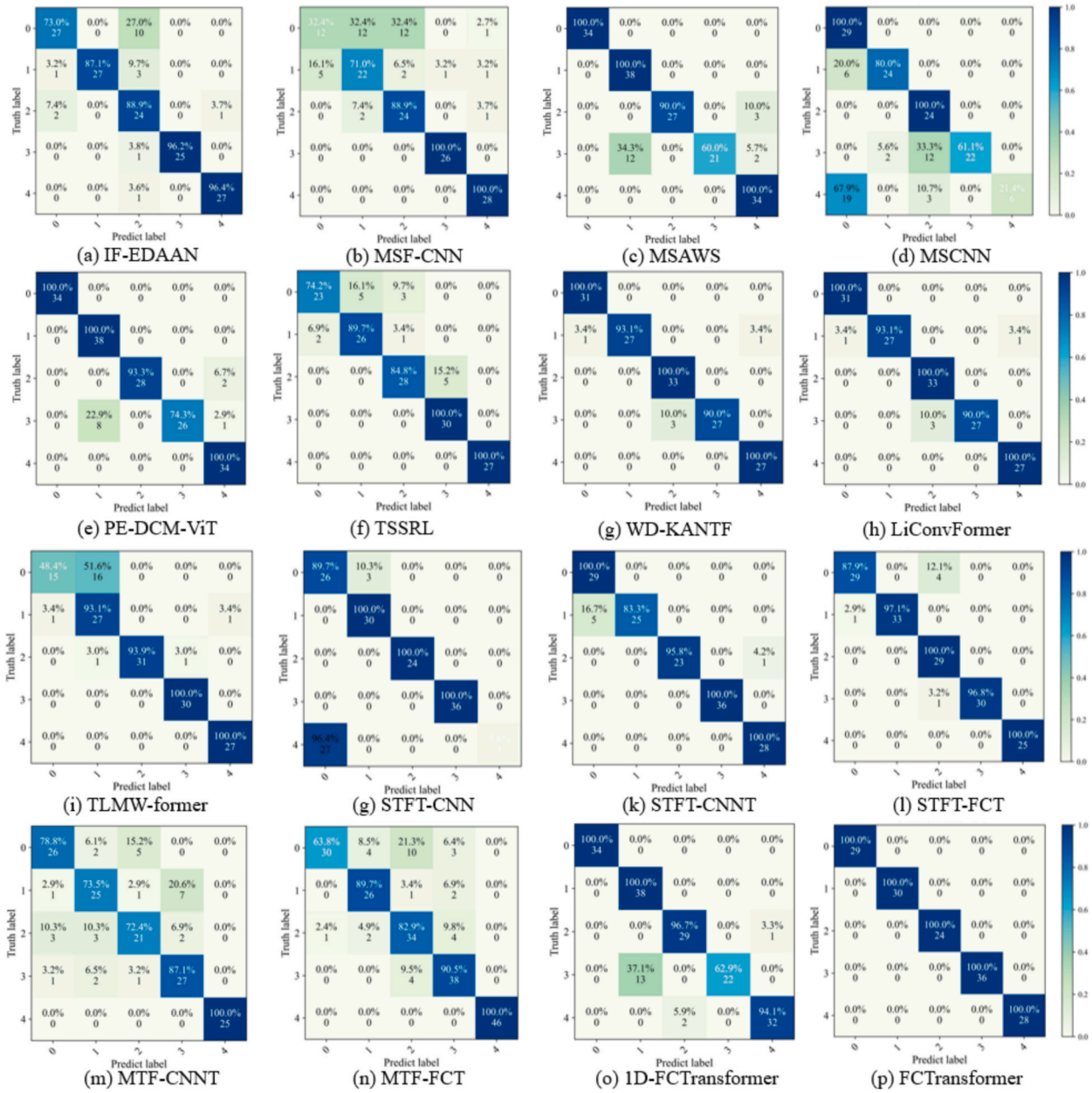


Fig. 14. Fault diagnosis visualization results for normal sample conditions.

feature extraction, enhancing the model's diagnostic performance under small sample conditions.

To further exhibit the advantages of VAETransformer and Fourier convolution, we conducted ablation experiments on the latent dimension of VAETransformer and the sine/cosine frequency parameters of Fourier convolution. Specifically, we selected 30/30/90 from the PBD dataset for experimentation. The impact of latent dimension on model performance was illustrated in Fig. 23, the left y-axis showed the impact of different latent dimensions on the model's diagnostic performance, the right y-axis displayed the results for Transformer and the proposed method in terms of floating-point operations. In Fig. 23, the model's fault diagnosis accuracy improved with increasing latent dimensions, and the model achieved the highest accuracy at latent $= 10$. However, when latent ≥ 10, the model's fault diagnosis accuracy decreased with increasing latent dimension. This was because the growing number of model parameters led to overfitting. In addition, the right y-axis indicated that the models of different latent dimensions had much lower floating-point operations than Transformer, this was because the proposed method utilized low-rank matrices, effectively reducing the number of model parameters.

Fig. 24 presented the impact of frequency parameters for the sine and cosine functions, the left y-axis displayed the model accuracy results, and the right y-axis showed the comparison of floating-point operations. Num $= 3$ represented the specific frequency setting for

Table 6
Fault diagnosis results of SEU dataset under small sample conditions.

Methods	Training set/ validation set/ test set							
	10/10/30		20/20/40		30/30/90		40/40/120	
	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall
IF-EDAAN	43.79%	40.43%	49.06%	47.83%	61.00%	60.99%	68.98%	67.84%
MSF-CNN	51.95%	51.69%	52.63%	53.14%	65.06%	64.95%	71.01%	68.12%
MSAWS	48.52%	48.63%	51.01%	50.67%	60.48%	59.32%	70.28%	69.79%
MSCNN	61.35%	59.60%	72.48%	70.74%	75.77%	75.63%	80.08%	80.34%
PE-DCM-ViT	42.76%	42.44%	52.80%	51.92%	60.24%	59.23%	68.58%	68.04%
TSSRL	51.33%	50.49%	54.00%	53.78%	57.11%	57.61%	71.50%	70.29%
WD-KANTF	59.33%	56.10%	71.24%	70.25%	82.44%	82.49%	85.67%	85.63%
LiConvFormer	54.00%	50.91%	63.00%	59.91%	63.33%	62.87%	75.33%	75.01%
TLMW-former	59.33%	61.39%	72.33%	72.09%	73.78%	73.69%	81.17%	80.73%
STFT-CNN	38.03%	37.31%	45.87%	45.05%	49.14%	49.09%	58.31%	56.23%
STFT-CNNT	38.66%	38.35%	46.39%	46.01%	50.76%	49.09%	80.01%	79.68%
SFFT-FCT	52.60%	52.04%	57.32%	55.88%	72.71%	72.44%	80.61%	80.43%
MTF-CNNT	53.12%	52.71%	64.77%	64.42%	65.66%	65.34%	73.01%	73.07%
MTF-FCT	58.46%	57.63%	74.82%	74.37%	75.06%	74.23%	83.62%	82.86%
1D-FCTransformer	61.33%	59.88%	68.04%	68.57%	74.57%	75.01%	83.67%	83.13%
FCTransformer	93.64%	93.20%	95.06%	94.96%	96.94%	96.64%	98.79%	98.70%
	Macro-F1	F1	Macro-F1	F1	Macro-F1	F1	Macro-F1	F1
IF-EDAAN	41.55%	43.00%	51.17%	48.42%	61.09%	61.54%	67.43%	66.16%
MSF-CNN	52.50%	48.96%	53.86%	53.36%	65.38%	65.00%	68.81%	67.30%
MSAWS	49.34%	50.00%	49.45%	50.56%	60.88%	59.61%	68.49%	68.61%
MSCNN	63.04%	60.97%	71.14%	72.97%	75.81%	74.68%	79.53%	81.75%
PE-DCM-ViT	50.56%	44.61%	49.45%	50.56%	58.44%	60.12%	69.33%	71.55%
TSSRL	50.92%	52.83%	53.18%	53.78%	56.78%	57.44%	71.66%	70.15%
WD-KANTF	57.06%	58.85%	71.15%	70.87%	80.87%	80.79%	83.48%	85.71%
LiConvFormer	52.31%	53.92%	60.62%	62.60%	58.23%	59.55%	74.51%	74.45%
TLMW-former	59.55%	60.79%	75.19%	73.98%	73.42%	73.82%	81.15%	81.23%
STFT-CNN	39.34%	40.03%	44.68%	43.33%	49.16%	47.21%	59.67%	61.05%
STFT-CNNT	42.23%	41.22%	45.21%	46.64%	48.43%	49.31%	79.19%	78.69%
SFFT-FCT	51.92%	50.00%	58.85%	58.09%	71.05%	70.28%	78.87%	80.95%
MTF-CNNT	52.58%	51.01%	64.02%	64.48%	66.04%	63.60%	71.52%	74.58%
MTF-FCT	57.80%	57.55%	76.19%	75.86%	75.58%	76.40%	84.58%	83.52%
1D-FCTransformer	61.11%	61.33%	67.74%	68.04%	73.05%	74.55%	84.81%	84.24%
FCTransformer	93.33%	92.30%	94.58%	94.09%	96.54%	96.37%	98.52%	98.54%

the sine and cosine functions, denoted as $\pi, 2\pi, 3\pi$. KANT used KAN to replace Fourier convolution to validate the linear approximation capability. The model achieved the highest diagnostic accuracy on both the PBD and SEU datasets at Num = 5, which had values of 97.79% and 98.79%, respectively. However, the accuracy gradually decreased on both datasets for Num $\geq$ 5, this indicated that the frequency of the sine and cosine functions significantly impacted the model's fault diagnosis performance. Excessively high frequency settings resulted in the model learning redundant information, which reduced fault diagnosis accuracy. Furthermore, the diagnostic accuracy of the proposed method was comparable to that of KANT at Num = 1. As the frequency parameter of sine and cosine functions increased, the model's diagnostic accuracy also improved, which further demonstrated the linear approximation capability of Fourier convolution.

The splitting ratio of training/validation/test sets affects the fault diagnosis performance. Therefore, this section sets different splitting ratios to validate the performance of the proposed method. Specifically, we selected 100 samples from the PBD and SUE datasets for the experiment, and the results were shown in Fig. 25, the horizontal axis indicated the dataset splitting ratio. For example, 60/10/30 meant that 60% of the data was used for model training, 10% for model validation, and 30% for model testing. Fig. 25 showed that the proposed method achieved the highest fault diagnosis accuracy at the validation set ratio of 0.2 for both PBD and SEU datasets. This indicated that the proposed method could fully utilize the training set for learning, while effectively adjusting the parameters through the validation set to achieve optimal fault diagnosis performance. The fault diagnosis accuracy of the proposed method was increasing with the number of validation samples at normal sample conditions. This was because more validation samples provided richer information, helping the model adjust the parameters more accurately. However, when the validation set was 25%, information leakage caused the fault diagnosis accuracy to decrease. Especially, it was worth noting that the impact of validation set size was more dramatic under small sample conditions. When the validation set was 25%, the fault diagnosis accuracy of the proposed method decreased by 5.22% and 3.27%, respectively.

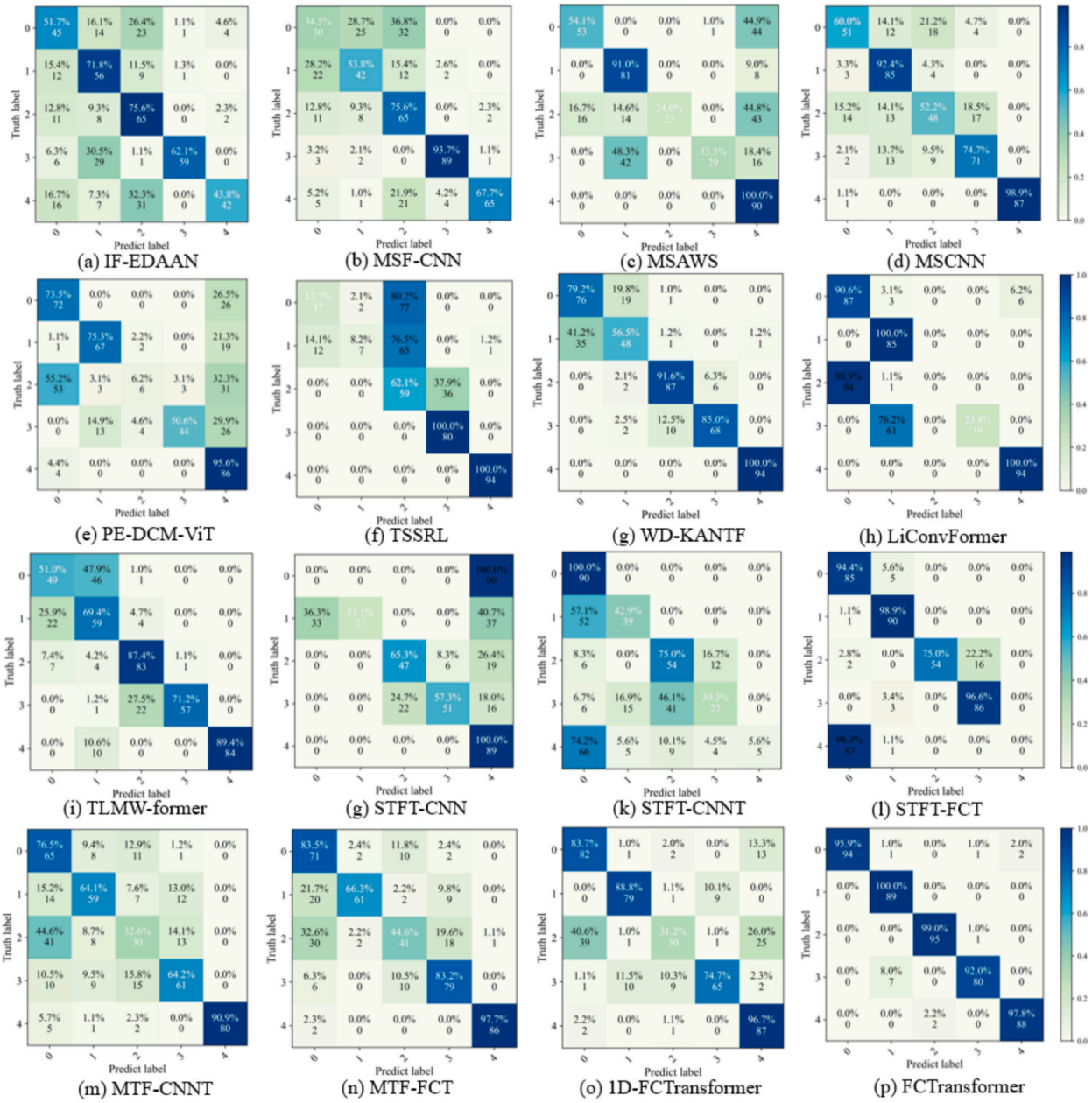


Fig. 15. Fault diagnosis visualisation results for small sample conditions.

4.10. Benchmark experiments based on the CWRU dataset

4.10.1. Dataset description

The CWRU dataset came from the bearing test platform at Case Western Reserve University. The test platform consisted of the drive-end motor, drive shaft, load, and variable frequency drive. The bearing status included normal, inner ring failure, outer ring failure, and rolling element failure. Each failure category was further divided into three severity levels: 0.007, 0.014, and 0.021. The CWRU dataset collected vibration signals at 0HP, 1HP, 2HP, and 3HP. Each vibration signal included data from the x-axis, y-axis, and z-axis, making this dataset suitable for multi-sensor data fusion research.

4.10.2. Experimental results

To validate the superiority of FCTransformer, its fault diagnosis results were compared fairly with those of published research methods. Specifically, MBSDCN [42] was a multi-branch multi-scale dynamic convolutional network was used for fault diagnosis of

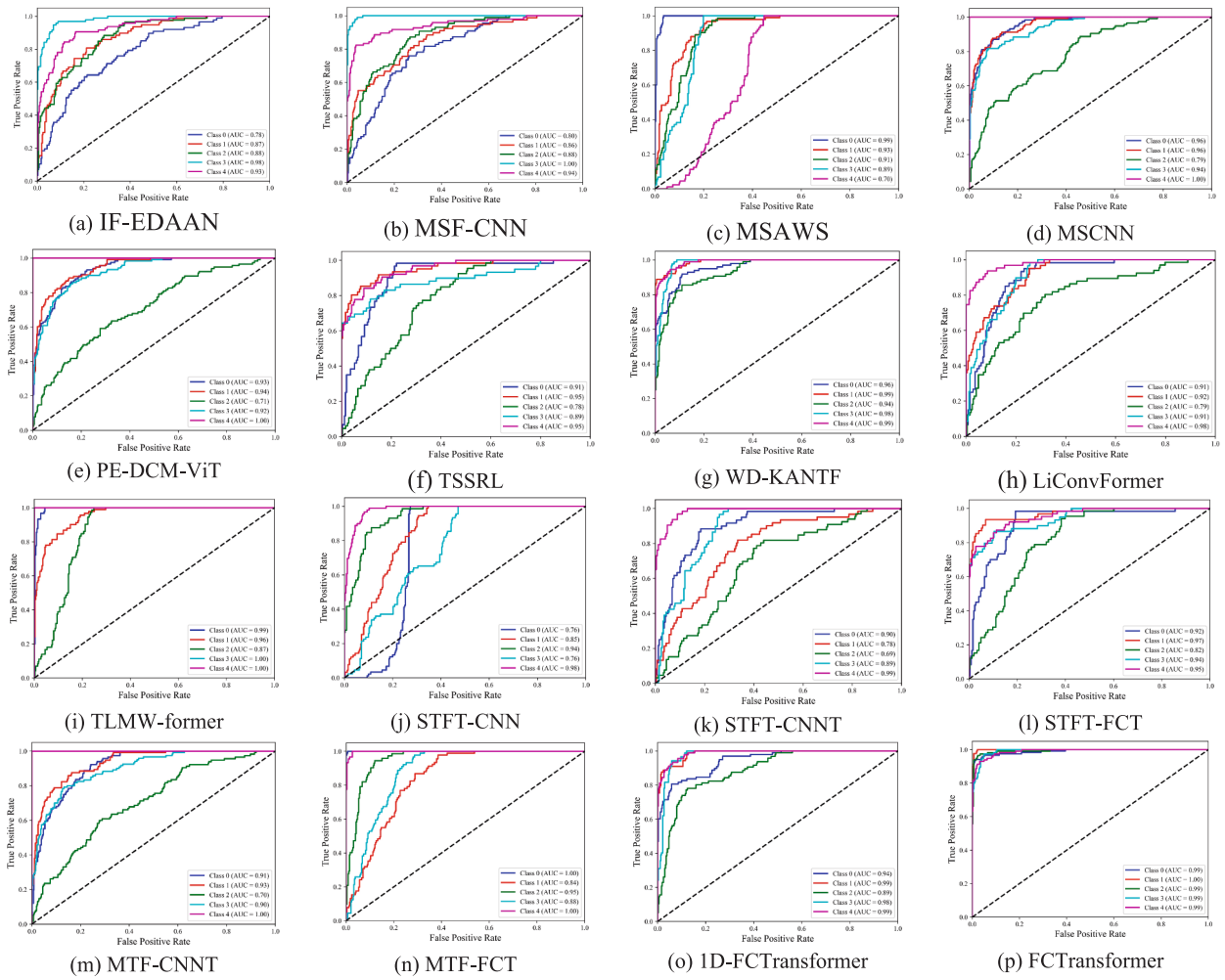


Fig. 16. The results of ROC curves for different methods on the PBD dataset.

rotating machinery under small sample conditions; WD-KANTF [39] was a fault diagnosis method based on wavelet denoising and KANTransfer; C-ECAFormer [30] was a lightweight model for fault diagnosis of rotating machinery; MSDFM-ASTRSB [43] was a multi-sensor data fusion module with asymmetric soft threshold residual shrinkage blocks for rolling bearing fault diagnosis.

Table 8 showed the comparison results with published methods, where the experimental results of the comparison methods were all sourced from the corresponding references. In Table 8, SNR (signal-to-noise ratio) referred to the addition of Gaussian white noise. In Table 8, the diagnostic accuracy of FCTransformer was lower than that of C-ECAFormer only under the 10/100 condition. In the other ten small-sample tasks, the diagnostic accuracy of FCTransformer was higher than that of the comparison methods. Especially under the small sample 50/50/100 condition, the diagnostic accuracy of FCTransformer was 100%. Furthermore, the average fault diagnosis accuracy of FCTransformer was 95.61%, which indicated that FCTransformer had more stable fault diagnosis performance. Compared to WD-KANTF, the average diagnostic accuracy of FCTransformer improved by 3.13%, which indicated that multi-sensor data fusion could obtain more fault information and enhance the diagnostic accuracy. Compared to MSDFM-ASTRSB, the average diagnostic accuracy of FCTransformer improved by 2.26%, which indicated that using VAETransformer for global dependency modelling could capture global fault features and enhance the fault diagnosis performance.

5. Conclusion

In this paper, an intelligent fault diagnosis method based on Fourier convolution and transformer (FCTransformer) is proposed to achieve more precise and intelligent diagnostics for rotating machinery. Specifically, a multi-sensor fusion strategy is designed to generate a multi-source data fusion dataset through dimensionality reduction and continuous wavelet transform, this dataset contains fault information from different locations of rotating machinery, effectively reducing the loss of fault information. Subsequently, a Fourier convolution is designed, which focuses on local periodic frequency-domain features through the sine and cosine convolution operators. Finally, a regularized VAETransformer module is proposed, which performs global dependency modeling by the variational

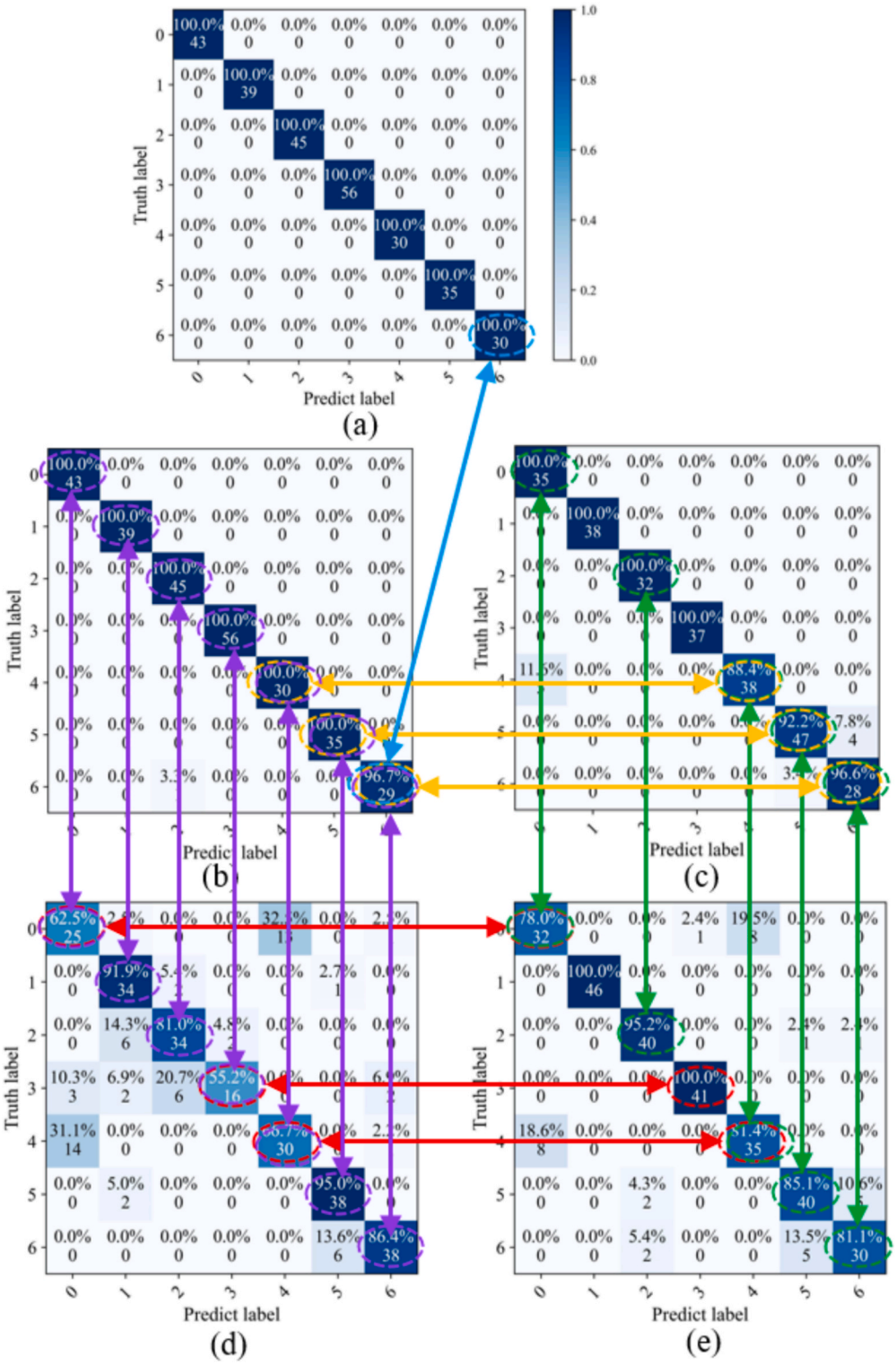


Fig. 17. Multi-sensor fusion fault diagnosis results for PBD dataset.

autoencoder attention mechanism. On the PBD dataset, the proposed method achieves an average fault diagnosis accuracy of 99.58%, which is superior to the comparison methods. On the SEU dataset, the proposed method achieves diagnostic accuracy exceeding 90.0%, which indicates the superior fault diagnosis performance.

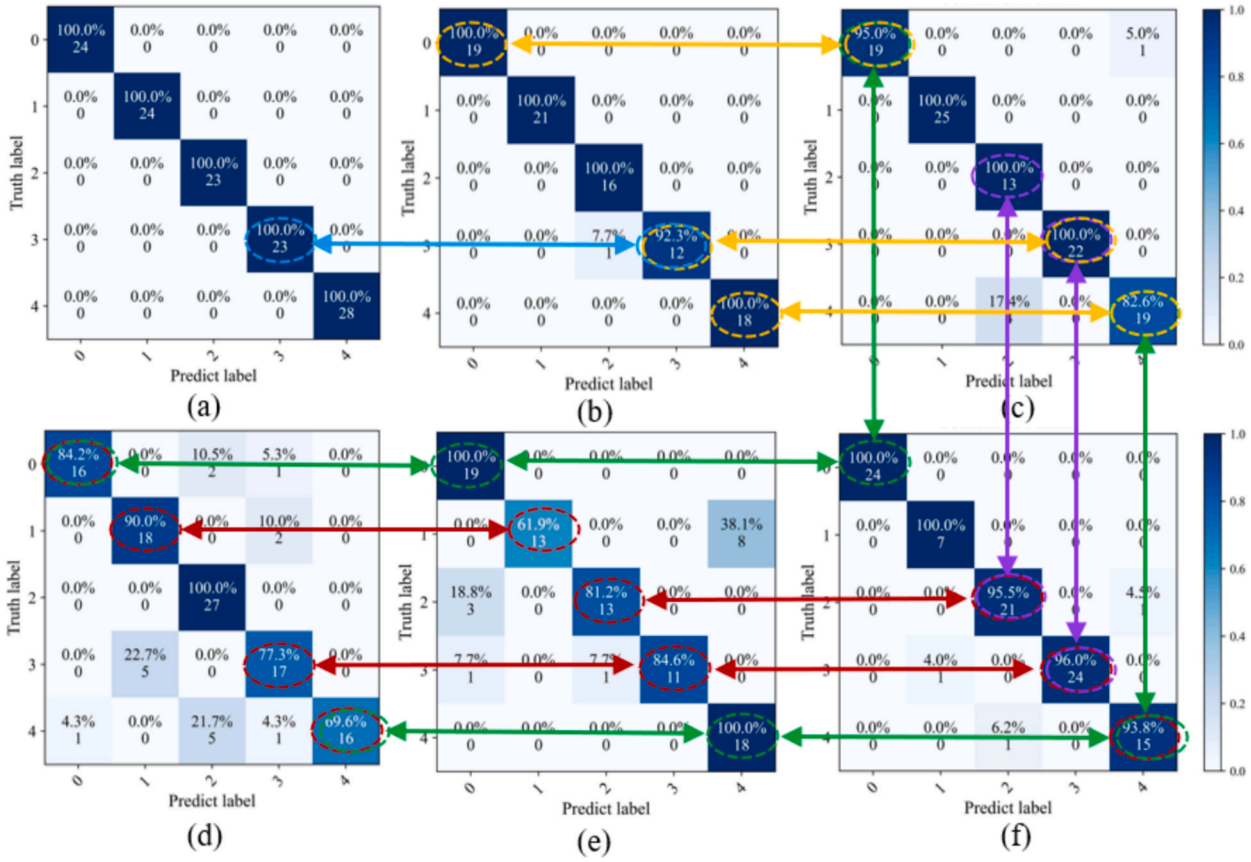


Fig. 18. Multi-sensor fusion fault diagnosis results of SEU dataset.

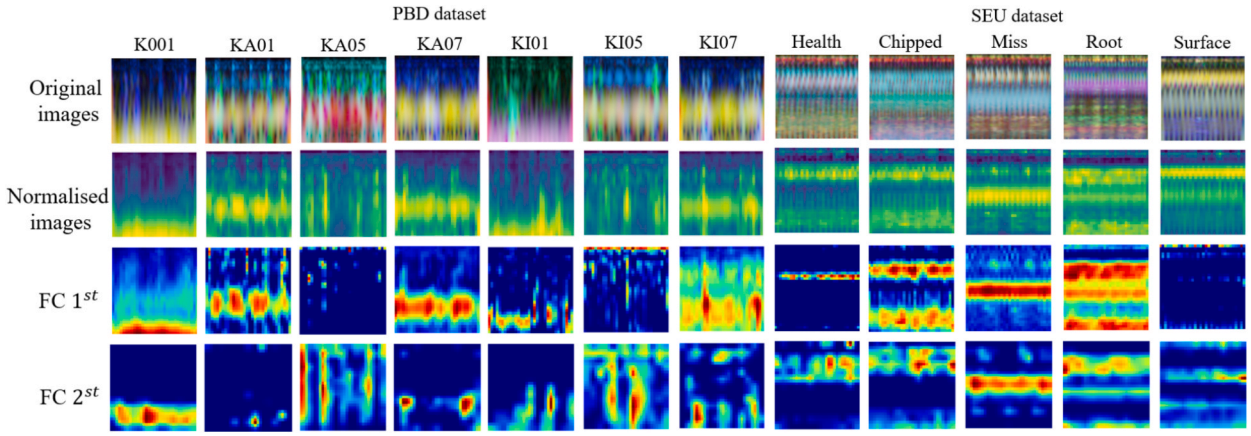


Fig. 19. Grad-CAM feature learning visualization result.

However, this study also presents several points of concern. (1) The multi-fusion strategy designed exhibits strong dependency on input data quality. Missing data or sample imbalance in the fused data can affect fault diagnosis results. (2) The data used in the experiments are collected under identical operating conditions. However, in real industrial environments, rotating machinery often operates under variable conditions and noisy conditions.

To address the above limitations, we will consider to use data augmentation techniques and generative adversarial networks in the design of fusion strategies, aiming to resolve information loss caused by data scarcity or sample imbalance. Furthermore, to better simulate industrial environments, we will introduce transfer learning strategies and noise-resistant techniques into the proposed method.

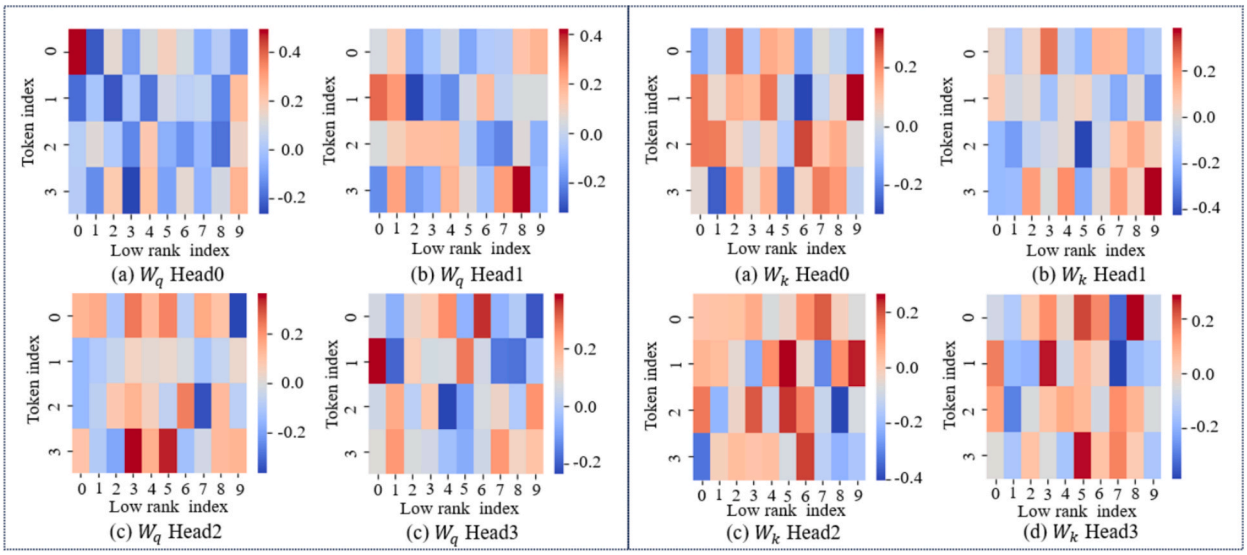


Fig. 20. Visualization of the weight matrix for the query and key matrices.

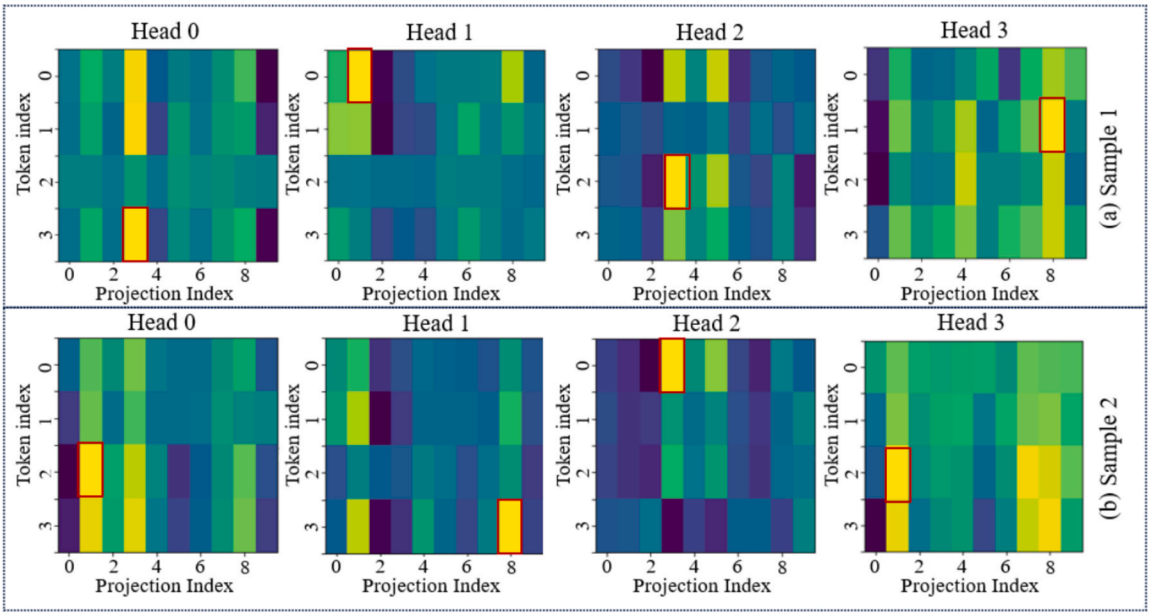


Fig. 21. Visualization results of the score values for the variational auto-coding attention mechanism.

Table 7
Ablation analysis of different modules.

A1	A2	A3	Training /validation /test (PBD)				Training /validation /test (SEU)			
			10/10/30	20/20/40	30/30/90	40/40/120	10/10/30	20/20/40	30/30/90	40/40/120
√	√	√	83.09%	87.52%	94.82%	97.79%	93.64%	95.06%	96.94%	98.79%
×	√	√	65.83%	70.62%	77.56%	85.54%	73.05%	77.27%	81.10%	86.64%
√	×	√	69.50%	77.38%	83.61%	88.64%	77.11%	82.54%	84.13%	91.43%
√	√	×	74.28%	80.12%	87.00%	92.18%	80.06%	85.39%	86.96%	94.25%

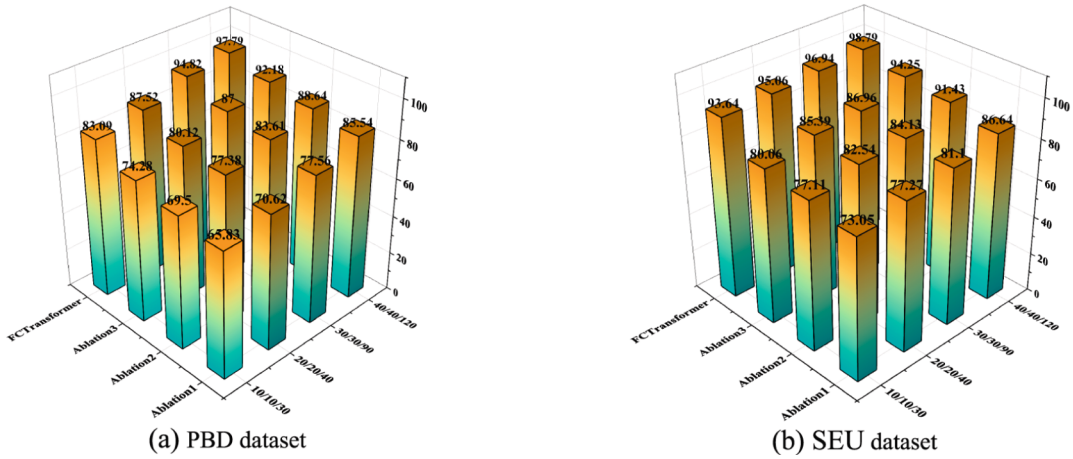


Fig. 22. Ablation experiment results for different modules.

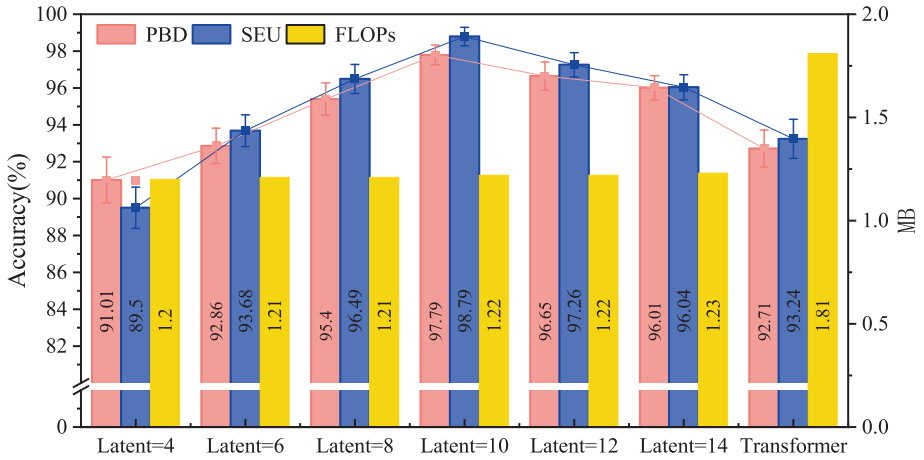


Fig. 23. Ablation experiment results for latent dimensions of VAETransformer.

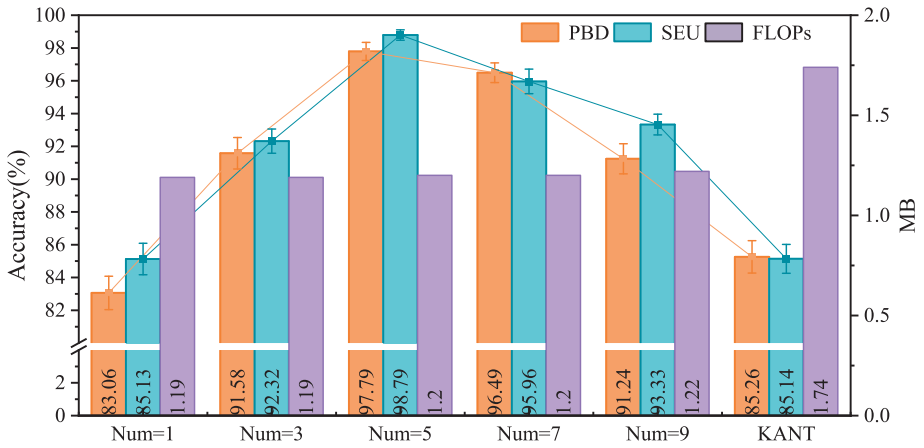


Fig. 24. Ablation experiment results for the frequency of sine and cosine functions.

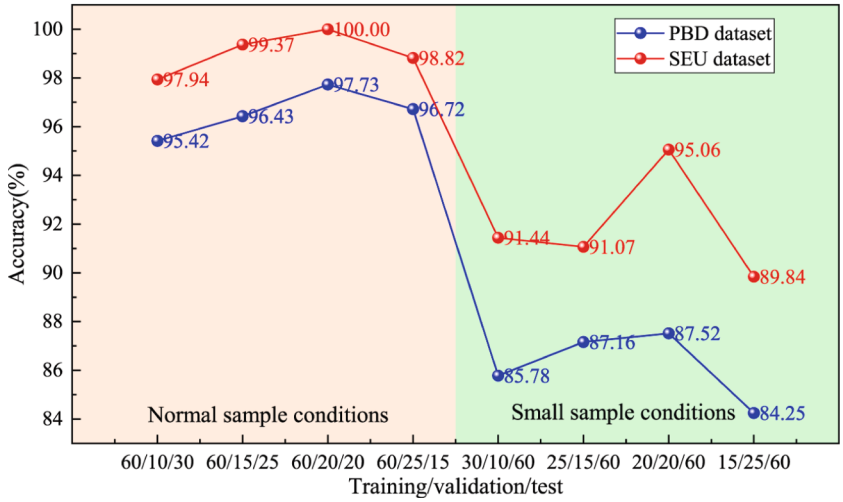


Fig. 25. Fault diagnosis results for different training/validation/test sets.

Table 8
Comparison results between FCTransformer and other published methods for CWRU dataset.

Condition	Training/Validation/Test Training/Test	MBSDCN	WD-KANTF	C-ECAFormer	MSDFM-ASTRSB	FCTransformer
0HP	10/10/30	–	86.67 ± 1.14%	–	–	91.12 ± 0.14%
SNR = 2 dB	20/20/60	–	93.50 ± 0.12%	–	–	96.64 ± 0.73%
	30/30/90	–	96.47 ± 0.22%	–	–	98.26 ± 0.34%
0HP	10/100	–	–	87.07%	–	85.84 ± 1.02%
SNR = −5dB	50/100	–		97.06%	–	97.32 ± 0.08%
1HP	5/5/100	–	–	–	81.67 ± 4.06%	89.88 ± 0.70%
SNR = 15 dB	10/10/100	–	–	–	97.47 ± 2.46%	98.06 ± 0.16%
	25/25/100	–	–	–	99.58 ± 0.42%	99.77 ± 0.23%
	50/50/100	–	–	–	99.92 ± 0.08%	100.0 ± 0.00%
3HP	20/40	91.15%	–	–	–	96.62 ± 1.17%
SNR = −3dB	40/80	92.97%	–	–	–	98.23 ± 0.21%

CRediT authorship contribution statement

Yazhou Zhang: Writing – original draft, Validation, Resources, Methodology, Funding acquisition. **Xiaoqiang Zhao:** Writing – review & editing, Supervision, Funding acquisition, Formal analysis. **Zhenrui Peng:** Supervision, Funding acquisition. **Yongyong Hui:** Supervision, Funding acquisition, Data curation. **Rongrong Xu:** Validation, Conceptualization.

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.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (No. 62263021), the College Industrial Support Project of Gansu Province (2023CYZC-24), the Science and Technology Project of Gansu Province (24JRRA172), and the Outstanding Postgraduate Innovation Star Project of Gansu Provincial Department of Education (2025CXZX-491), the Gansu Provincial Basic Research Innovation Group of China (25JRRA058), the Central Government's Funds for Guiding Local Science and Technology Development of China (25ZYJA040)

Appendix

Proposition: the Kolmogorov-Arnold theorem states that for any function $f(x_1, x_2, ..., x_v)$, there exists a univariate continuous function $\varepsilon_{p,q}$, which satisfies the following formula:

$$f(x_1, x_2, ..., x_v) = \sum_{p=0}^{2v+1} \varepsilon_{p,q}\left(\sum_{q=1}^{v} \varepsilon_{p,q}(x_p)\right) \tag{1}$$

where v is the number of independent variables for the multivariate continuous function $f(\cdot)$, p and q denote the nodes of KAN, $p \in \{0, 1, 2, ..., 2v+1\}$, $q \in \{1, 2, 3, ..., n\}$.

To prove that KAN retains approximation after replacing univariate continuous functions with sine and cosine functions, it is necessary to demonstrate that for any given $\zeta > 0$, there exists a finite combination of sine and cosine functions $\phi_{p,q}(x_p)$, which satisfies the following formula:

$$\left| f(x_1, x_2, ..., x_v) - \sum\nolimits_{p=0}^{2v+1} \phi_{p,q}\left(\sum\nolimits_{q=1}^{v} \phi_{p,q}(x_p)\right) \right| < \zeta, \quad x_p \in R \tag{2}$$

Formula 2 demonstrates that KAN retains its approximation capability after replacing univariate continuous functions with sine and cosine functions. It is necessary to prove that the overall error can be controlled to arbitrarily small values using a finite combination of sine and cosine functions for each $\varepsilon_{p,q}$.

Certification: the proof process consists of the following two steps.

Step 1: Prove that the Fourier series of a continuous function with one element has approximation capability.

Supposed $F(x)$ is a continuous function defined on the interval $[-\pi, \pi]$. According to the Fourier series theory, the Fourier series of $F(x)$ can be expressed as follows:

$$F(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos(nx) + bn \sin(nx)) \tag{3}$$

where a_n and b_n are defined as follows:

$$\begin{cases} a_n = \dfrac{1}{\pi}\displaystyle\int_{-\pi}^{\pi} F(x)\cos(nx)dx, & n = 0, 1, 2, ... \\ b_n = \dfrac{1}{\pi}\displaystyle\int_{-\pi}^{\pi} F(x)\sin(nx)dx, & n = 1, 2, ... \end{cases} \tag{4}$$

According to the Dirichlet-Jordan theorem, if $F(x)$ is continuous and piecewise smooth on $[-\pi, \pi]$, the Fourier series converges uniformly to on $[-\pi, \pi]$. Thus, there exists a positive integer N_1 for any given $\zeta > 0$. When $N \geq N_1$, this satisfies the following formula:

$$\left| F(x) - \left(\frac{a_0}{2} + \sum\nolimits_{n=1}^{N} (a_n \cos(nx) + bn \sin(nx))\right) \right| < \zeta, \quad x \in [-\pi, \pi] \tag{5}$$

From Formula (5), it can be seen that a continuous function of one variable can be approximated with arbitrary precision by a finite linear combination of sine and cosine functions.

Step 2: Prove that KAN has approximation capability after replacing a univariate continuous function with sine and cosine functions.

Based on the Fourier series approximation proven in Step 1, there exists a finite combination $\zeta > 0$ of sine and cosine functions for any given $\phi_{p,q}(x_p)$, which satisfies the following formula:

$$\left|\varepsilon_{p,q}(x_p) - \phi_{p,q}(x_p)\right| < \zeta_1, \quad x_p \in R \tag{6}$$

where $\phi_{p,q}(x_p) = \frac{a_{0p,q}}{2} + \sum_{k=1}^{N_{q,p}} (a_{k,p,q}\cos(kx_p) + b_{k,p,q}\sin(kx_p))$, $a_{k,p,q}$ and $b_{k,p,q}$ are the Fourier coefficients.

The approximation formula for the sine and cosine functions is obtained by substituting $\phi_{p,q}(x_p)$ in Formula (1) with $\varepsilon_{p,q}$, which is as follows:

$$\widehat{f}(x_1, x_2, ..., x_v) = \sum_{p=0}^{2v+1} \phi_{p,q}\left(\sum_{q=1}^{v} \phi_{p,q}(x_p)\right) \tag{7}$$

Estimate the error of the inner summation of $f(x_1, x_2, ..., x_v)$ and $\widehat{f}(x_1, x_2, ..., x_v)$. Specifically, we calculate the error between $\sum_{q=1}^{v} \varepsilon_{p,q}(x_p)$ and $\sum_{q=1}^{v} \phi_{p,q}(x_p)$, which is described by:

$$\left| \sum_{q=1}^{v} \varepsilon_{p,q}(x_p) - \sum_{q=1}^{v} \phi_{p,q}(x_p) \right| = \left| \sum_{q=1}^{v} (\varepsilon_{p,q}(x_p) - \phi_{p,q}(x_p)) \right| \leq \sum_{q=1}^{v} \left|\varepsilon_{p,q}(x_p) - \phi_{p,q}(x_p)\right| \tag{8}$$

where, $\left|\varepsilon_{p,q}(x_p)-\phi_{p,q}(x_p)\right|<\zeta_1$ can be obtained from Formula (6). Thus, Formula (8) can be expressed as $\left|\sum_{q=1}^{v}\varepsilon_{p,q}(x_p)-\sum_{q=1}^{v}\phi_{p,q}(x_p)\right|<v\zeta_1$.

Estimate the error of the the outer summation of $f(x_1,x_2,\ldots,x_v)$ and $\widehat{f}(x_1,x_2,\ldots,x_v)$. Specifically, since $\varepsilon_{p,q}(x_p)$ is Lipschitz continuous, there exists the following description:

$$\left|\varepsilon_{0,q}\left(\sum_{q=1}^{v}\varepsilon_{p,q}(x_p)\right)-\phi_{0,q}\left(\sum_{q=1}^{v}\phi_{p,q}(x_p)\right)\right|\le\left|\varepsilon_{0,q}\left(\sum_{q=1}^{v}\varepsilon_{p,q}(x_p)\right)-\varepsilon_{0,q}\left(\sum_{q=1}^{v}\phi_{p,q}(x_p)\right)\right|+\left|\varepsilon_{0,q}\left(\sum_{q=1}^{v}\phi_{p,q}(x_p)\right)-\phi_{0,q}\left(\sum_{q=1}^{v}\phi_{p,q}(x_p)\right)\right| \tag{9}$$

The Lipschitz property obtains the following formula for $\left|\varepsilon_{0,q}\left(\sum_{q=1}^{v}\varepsilon_{p,q}(x_p)\right)-\varepsilon_{0,q}\left(\sum_{q=1}^{v}\phi_{p,q}(x_p)\right)\right|$.

$$\left|\varepsilon_{0,q}\left(\sum_{q=1}^{v}\varepsilon_{p,q}(x_p)\right)-\varepsilon_{0,q}\left(\sum_{q=1}^{v}\phi_{p,q}(x_p)\right)\right|\le L_P\left|\sum_{q=1}^{v}\varepsilon_{p,q}(x_p)-\sum_{q=1}^{v}\phi_{p,q}(x_p)\right|<L_P v\zeta_1 \tag{10}$$

where L_p is a constant, and $L_p>0$.

For $\left|\varepsilon_{0,q}\left(\sum_{q=1}^{v}\phi_{p,q}(x_p)\right)-\phi_{0,q}\left(\sum_{q=1}^{v}\phi_{p,q}(x_p)\right)\right|$, since $\phi_{0,q}$ is a Fourier series approximation of $\varepsilon_{0,q}$, when the number of Fourier terms $N_{0,p}$ is sufficiently large for any given $\zeta_2>0$, the following formula is satisfied as follows:

$$\left|\varepsilon_{0,q}\left(\sum_{q=1}^{v}\phi_{p,q}(x_p)\right)-\phi_{0,q}\left(\sum_{q=1}^{v}\phi_{p,q}(x_p)\right)\right|<\zeta_2 \tag{11}$$

From the above derivation, it can be seen that the errors in $f(x_1,x_2,\ldots,x_v)$ and $\widehat{f}(x_1,x_2,\ldots,x_v)$ can be described as follows:

$$\begin{aligned}\left|f(x_1,x_2,\ldots,x_v)-\widehat{f}(x_1,x_2,\ldots,x_v)\right|&=\left|\sum_{p=1}^{2v+1}\left[\varepsilon_{0,q}\left(\sum_{q=1}^{v}\varepsilon_{p,q}(x_p)\right)-\phi_{0,q}\left(\sum_{q=1}^{v}\phi_{p,q}(x_p)\right)\right]\right|\\&\le\sum_{p=1}^{2v+1}\left|\varepsilon_{0,q}\left(\sum_{q=1}^{v}\varepsilon_{p,q}(x_p)\right)-\phi_{0,q}\left(\sum_{q=1}^{v}\phi_{p,q}(x_p)\right)\right|\\&<\sum_{p=1}^{2v+1}\left(L_p v\zeta_1+\zeta_2\right)\end{aligned} \tag{12}$$

From Formula (12), it can be seen that $\sum_{p=1}^{2v+1}\left(L_p v\zeta_1+\zeta_2\right)<\zeta$ is satisfied for any given $\zeta>0$ when ζ_1 and ζ_2 are sufficiently small.

In conclusion, after replacing the univariate functions in KAN with cosine and sine functions, its universal approximation is still valid.

Data availability

Data will be made available on request.

References

[1] M. Niu, H. Jiang, H. Shao, Dynamic weighted adversarial domain adaptation network with sparse representation denoising module for rotating machinery fault diagnosis, Eng. Appl. Artif. Intel. 142 (2025) 109963.
[2] J. Zhang, K. Zhang, Y. An, H. Luo, S. Yin, An integrated multitasking intelligent bearing fault diagnosis scheme based on representation learning under imbalanced sample condition, IEEE Trans. Neural Networks Learn. Syst. 35 (2024) 6231–6242.
[3] C. Yang, et al., Fast and stable fault diagnosis method for composite fault of subsea production system, Mech. Syst. Sig. Process. 226 (2025) 112373.
[4] Y. Zhang, X. Zhao, R. Xu, Feature and Joint distribution Migration Alignment Method for Cross-domain Fault Diagnosis of Rotating Machinery, IEEE Trans. Instrum. Meas. 74 (2025) 3525115.
[5] Q. Qian, Q. Wen, R. Tang, Y. Qin, DG-Softmax: a new domain generalization intelligent fault diagnosis method for planetary gearboxes, Reliab. Eng. Syst. Saf. (2025) 111057.
[6] X. Wang, H. Jiang, M. Mu, Y. Dong, A trackable multi-domain collaborative generative adversarial network for rotating machinery fault diagnosis, Mech. Syst. Sig. Process. 224 (2025) 111950.
[7] Z. Chen, J. Ji, W. Yu, Q. Ni, G. Lu, X. Chang, A multi-scale graph convolutional network with contrastive-learning enhanced self-attention pooling for intelligent fault diagnosis of gearbox, Measurement 230 (2024) 114497.
[8] F. Chen, Z. Zhao, X. Hu, D. Liu, X. Yin, J. Yang, A nonlinear dynamics method using multi-sensor signal fusion for fault diagnosis of rotating machinery, Adv. Eng. Inf. 65 (2025) 103190.
[9] Z. Zhao, Y. Jiao, A fault diagnosis method for rotating machinery based on CNN with mixed information, IEEE Trans. Ind. Inf. 19 (8) (2022) 9091–9101.
[10] Z. Li, H. Jiang, Y. Dong, A convolutional-transformer reinforcement learning agent for rotating machinery fault diagnosis, Expert Syst. Appl. 271 (2025) 126669.
[11] Q. Zhou, J. Tang, An interpretable parallel spatial CNN-LSTM architecture for fault diagnosis in rotating machinery, IEEE Internet Things J. 11 (19) (2024) 31730–31744.

[12] Y. Luo, M. Wang, L. Luo, Z. Liu, J. Zhao, Optimized LSTM-BiTCN parallel network model for anomalous sound detection in rotating machinery, Meas. Sci. Technol. 36 (4) (2025) 046110.
[13] C. He, H. Shi, X. Liu, J. Li, Interpretable physics-informed domain adaptation paradigm for cross-machine transfer diagnosis, Knowl.-Based Syst. 288 (2024) 111499.
[14] X. Zhao, Y. Zhang, An intelligent diagnosis method of rolling bearing based on multi-scale residual shrinkage convolutional neural network, Meas. Sci. Technol. 33 (8) (2022) 085103.
[15] F. Lu, Q. Tong, X. Jiang, X. Du, J. Xu, J. Huo, Prior knowledge embedding convolutional autoencoder: a single-source domain generalized fault diagnosis framework under small samples, Comput. Ind. 164 (2025) 104169.
[16] Y. Zhang, X. Zhao, H. Liang, P. Chen, Multiscale dilated convolution and swin-transformer for small sample gearbox fault diagnosis, Appl. Intell. 54 (17) (2024) 7716–7732.
[17] Y. Xiao, H. Shao, J. Wang, S. Yan, B. Liu, Bayesian variational transformer: a generalizable model for rotating machinery fault diagnosis, Mech. Syst. Sig. Process. 207 (2024) 110936.
[18] K. Zhou, C. Yang, J. Liu, Q. Xu, Deep graph feature learning-based diagnosis approach for rotating machinery using multi-sensor data, J. Intell. Manuf. 34 (4) (2023) 1965–1974.
[19] D. Wang, Y. Li, L. Jia, Y. Song, Y. Liu, Novel three-stage feature fusion method of multimodal data for bearing fault diagnosis, IEEE Trans. Instrum. Meas. 70 (2021) 1–10.
[20] Z. Yang, G. Li, G. Xue, B. He, Y. Song, X. Li, A novel multi-sensor local and global feature fusion architecture based on multi-sensor sparse transformer for intelligent fault diagnosis, Mech. Syst. Sig. Process. 224 (2025) 112188.
[21] D. Sun, Y. Li, S. Jia, S. Gao, K. Noman, K. Eliker, Physical knowledge-driven feature fusion and reconstruction network for fault diagnosis with incomplete multisource data, Mech. Syst. Sig. Process. 225 (2025) 112222.
[22] Y. Dong, H. Jiang, M. Mu, X. Wang, Multi-sensor data fusion-enabled lightweight convolutional double regularization contrast transformer for aerospace bearing small samples fault diagnosis, Adv. Eng. Inf. 62 (2024) 102573.
[23] H. Shao, J. Lin, L. Zhang, D. Galar, U. Kumar, A novel approach of multisensory fusion to collaborative fault diagnosis in maintenance, Inf. Fusion 74 (2021) 65–76.
[24] W. Gong, Y. Wang, M. Zhang, E. Mihankhah, H. Chen, D. Wang, A fast anomaly diagnosis approach based on modified CNN and multisensor data fusion, IEEE Trans. Ind. Electron. 69 (12) (2021) 13636–13646.
[25] C. Lin, et al., IF-EDAAN: an information fusion-enhanced domain adaptation attention network for unsupervised transfer fault diagnosis, Mech. Syst. Sig. Process. 224 (2025) 112180.
[26] M. Ye, X. Yan, D. Jiang, L. Xiang, N. Chen, MIFDELN: a multi-sensor information fusion deep ensemble learning network for diagnosing bearing faults in noisy scenarios, Knowl.-Based Syst. 284 (2024) 111294.
[27] H. Liang, J. Cao, X. Zhao, Multi-scale dynamic adaptive residual network for fault diagnosis, Measurement 188 (2022) 110397.
[28] Y. Li, et al., Graph optimization algorithm enhanced by dual-scale spectral features with contrastive learning for robust bearing fault diagnosis, Knowl.-Based Syst. 315 (2025) 113275.
[29] S. Dong, Y. Meng, S. Yin, X. Liu, Tool wear state recognition study based on an MTF and a vision transformer with a Kolmogorov-Arnold network, Mech. Syst. Sig. Process. 228 (2025) 112473.
[30] J. Wang, H. Shao, S. Yan, B. Liu, C-ECAFormer: a new lightweight fault diagnosis framework towards heavy noise and small samples, Eng. Appl. Artif. Intel. 126 (2023) 107031.
[31] Z. Zhang, N. Shao, C. Gao, R. Miao, Q. Yang, J. Shao, Mixhead: breaking the low-rank bottleneck in multi-head attention language models, Knowl.-Based Syst. 240 (2022) 108075.
[32] J. Liu, et al., MGTN-DSI: a multi-sensor graph transfer network considering dual structural information for fault diagnosis under varying working conditions, Adv. Eng. Inf. 65 (2025) 103119.
[33] S. Shao, S. McAleer, R. Yan, P. Baldi, Highly accurate machine fault diagnosis using deep transfer learning, IEEE Trans. Ind. Inf. 15 (4) (2018) 2446–2455.
[34] T. Xie, X. Huang, S.-K. Choi, Intelligent mechanical fault diagnosis using multisensor fusion and convolution neural network, IEEE Trans. Ind. Inf. 18 (5) (2021) 3213–3223.
[35] X. Jiang, X. Li, Q. Wang, Q. Song, J. Liu, Z. Zhu, Multi-sensor data fusion-enabled semi-supervised optimal temperature-guided PCL framework for machinery fault diagnosis, Inf. Fusion 101 (2024) 102005.
[36] L. Xue, C. Lei, M. Jiao, J. Shi, J. Li, Rolling bearing fault diagnosis method based on self-calibrated coordinate attention mechanism and multi-scale convolutional neural network under small samples, IEEE Sens. J. 23 (9) (2023) 10206–10214.
[37] Y. Hui, K. Xu, P. Chen, X. Zhao, Rolling bearing fault diagnosis method based on PE-DCM and ViT, Meas. Sci. Technol. 35 (10) (2024) 105107.
[38] M. Hua, K. Yan, X. Li, A Transformer-based self-supervised learning model for fault diagnosis of air-conditioning systems with limited labeled data, Eng. Appl. Artif. Intel. 146 (2025) 110331.
[39] Y. Zhang, X. Zhao, Z. Peng, R. Xu, P. Chen, WD-KANTF: an interpretable intelligent fault diagnosis framework for rotating machinery under noise environments and small sample conditions, Adv. Eng. Inf. 66 (2025) 103452.
[40] S. Yan, H. Shao, J. Wang, X. Zheng, B. Liu, LiConvFormer: a lightweight fault diagnosis framework using separable multiscale convolution and broadcast self-attention, Expert Syst. Appl. 237 (2024) 121338.
[41] Y. Dong, H. Jiang, M. Mu, X. Wang, A trustworthy lightweight multi-expert wavelet transformer for rotating machinery fault diagnosis, Mech. Syst. Sig. Process. 235 (2025) 112945.
[42] H. Liang, J. Cao, X. Zhao, Multibranch and multiscale dynamic convolutional network for small sample fault diagnosis of rotating machinery, IEEE Sens. J. 23 (8) (2023) 8973–8988.
[43] H. Guo, X. Zhao, MSDFM-ASTRSB: a rolling bearing fault diagnosis method with limited samples, IEEE Trans. Instrum. Meas. 73 (2024) 3541714.

Advanced Engineering Informatics 66 (2025) 103452

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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

Full length article

WD-KANTF: An interpretable intelligent fault diagnosis framework for rotating machinery under noise environments and small sample conditions

Yazhou Zhang [a], Xiaoqiang Zhao [a,b,*], Zhenrui Peng [a,b], Rongrong Xu [a], Peng Chen [c]

[a] College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China
[b] Gansu Key Laboratory of Advanced Control of Industrial Processes, Lanzhou 730050, China
[c] College of Electrical and Electronic Engineering, Lanzhou Petrochemical University of Technology, Lanzhou 730050, China

ARTICLE INFO

Keywords:
Fault diagnosis
Rotating machinery
Small sample conditions
Noise environments
Wavelet transform
Interpretability

ABSTRACT

In industrial applications, it is crucial that fault diagnosis is performed on rotating machinery to ensure production safety and improve efficiency. However, due to insufficient fault data and interference from noise environments, current intelligent diagnostic models face the challenge of decreased diagnostic accuracy in practical applications. In addition, the lack of interpretability further undermines the reliability of fault diagnosis. To address these challenges, a fault diagnosis framework based on wavelet denoising and KANTransformer (WD-KANTF) is proposed to enhance both diagnostic accuracy and interpretability. Firstly, an adaptive wavelet denoising layer is designed, it consists of discrete wavelet transform (DWT), smoothed soft threshold and fusion strategy. DWT is used to map the time domain features to the wavelet domain, so as to more efficiently capture the low-frequency and high-frequency features. Meanwhile, the noise interference is filtered out by using the smoothed soft threshold to improve the robustness. Secondly, KANTransformer is developed, which enhances the nonlinear extraction capability by introducing a learnable activation function in the linear layer. Finally, to verify the effectiveness of WD-KANTF, the experiments are conducted on the bearing dataset and the gearbox dataset. The results demonstrate that WD-KANTF has superior diagnostic performance and robustness under strong noise environments and small sample conditions. Furthermore, the interpretability analysis of WD-KANTF is conducted by time-series gradient activation mapping (TGAM) and weight visualization techniques, which confirm that it can accurately extract critical fault features and provide reliable support for practical industrial applications.

1. Introduction

Currently, rotating machinery is widely used in the fields of energy, transport and aerospace. However, with the increasing complexity and automation of rotating machinery, the operating environment of its critical components such as rolling bearings and gearboxes has become more demanding. Failure types such as cracks, fatigue degradation, spalling and deformation are increasingly occurring [1–3]. If these failures are not recognised and repaired, they would directly threaten the overall reliability of rotating machinery and may even lead to major safety accidents. Therefore, health monitoring of critical components has become an indispensable condition to ensure the long-term stable and reliable operation of rotating machinery [4–6].

Fault diagnosis methods for rotating machinery can be categorised as model-based and data-driven methods. Model-based methods require the development of accurate mathematical models [7,8], this poses a great challenge to the increasingly complex and intelligent rotating machinery. In recent years, with the rapid development of sensors and storage technologies, data-driven methods are widely popular among scholars [9,10]. In particular, the rapid development of deep learning has brought great changes for fault diagnosis. For example, Liang et al [11] developed a multiscale dynamic adaptive residual network, this network utilised multi-scale techniques and attention mechanisms to construct convolutional layers, which could dynamically adjust the weights and enhance the feature extraction performance. Shi et al [12] proposed a transferable siamese network, which extracted fault features by 1D-convolutional layers, residual structure and attention mechanism. Gong et al [13] proposed an improved convolutional neural network for fault diagnosis, which used one-dimensional global average pooling to replace the classification layer, thus speeding up the training of the

* Corresponding author at: College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China.
E-mail address: xqzhao@lut.edu.cn (X. Zhao).

https://doi.org/10.1016/j.aei.2025.103452
Received 25 November 2024; Received in revised form 30 March 2025; Accepted 4 May 2025

network. Dong et al [14] proposed a framework for multiscale dynamically supervised comparison learning, which used multiscale techniques and channel space attention mechanism to enhance the feature learning capability. Wu et al [15] proposed a deep adversarial model for imbalance fault diagnosis of bearings, which used cost-sensitive losses to solve the class imbalance problem. Qian et al [16] proposed a new discrepancy metric function to achieve fault transfer diagnosis, and constructed the maximum mean square discrepancy to reduce the difference between source and target domains. He et al [17] proposed a wavelet domain adaptive network based on physical information, which increased the interpretability by introducing wavelet knowledge. Although the above methods have shown encouraging results of fault diagnosis in rotating machinery, it is usually difficult to obtain lots of fault samples in practical industries. Meanwhile, harsh environments such as strong corrosion and high pressures not only damage the equipment itself, but also make data acquisition extremely difficult. Therefore, training fault diagnosis models with superior performance under small sample conditions is a challenge that needs to be addressed urgently.

Currently, small sample fault diagnosis methods can be classified into three categories: data augmentation, few-shot learning, and transfer learning. Data augmentation generates more training samples to expand the small sample dataset by generative adversarial network (GAN). Few-shot learning allows a model to extract useful information from a small number of samples through *meta*-training and to satisfy new task requirements in the *meta*-testing phase. Transfer learning is done by learning fault information from source tasks and transferring the fault information into target tasks. For example, Chen et al [18] embedded a channel attention mechanism into GAN to generate higher quality missing data, which ultimately achieved fault diagnosis under small sample conditions. Liu et al [19] proposed a generalized agnostic *meta*-learning framework, the framework used weight-guided factors to optimize the *meta*-learning training strategy, which improved the generalization performance of the model under small sample conditions. Sun et al [20] designed a pseudo-label-guided adversarial network for dual classifier domains, which efficiently reduced the misclassification rate for the target domains under small sample conditions. Although the above methods achieve high fault diagnosis accuracy under small sample conditions, they also have the following drawbacks. For example, using GAN to generate samples in data augmentation is subject to training instability, resulting in the generated samples not accurately reflecting the real fault data. Few-shot learning with *meta*-learning would be limited in its generalization ability when facing new tasks, making it difficult to meet the requirements of new tasks. Transfer learning may suffer from overfitting in the target domain with insufficient samples, thus affecting the diagnostic accuracy in the target domain. In addition, since noise can blur fault features, which leads to difficulty for the above methods to adequately extract fault features under noise environments and small sample conditions.

Based on the above reasons, some scholars have introduced the attention mechanism to enhance the feature extraction capability of the model. For example, Liang et al [21] designed a multi-branch multiscale dynamic convolutional network based on the feature segmentation strategy, this network uses the feature segmentation strategy to obtain multiple sub-signal features and enhances the useful fault features through the channel reconstruction attention mechanism. Wang et al [22] proposed a lightweight model of C-ECAFormer, the model designed collaborative self-attention modules using an efficient attention mechanism, which facilitated the spatial interaction between local and global features. Sun et al [23] effectively extracted fault features in noise environments using adaptive hybrid pooling and parallel attention mechanism. However, the above methods are based on convolutional neural networks (CNNs), due to the 'black box' characteristic of CNNs, they lack interpretability and can hardly reveal the internal reasoning process.

In order to improve the interpretability of the model, scholars have tried to introduce physics knowledge into deep learning. For example, He et al [17] proposed a physically based wavelet domain adaptive network, which was updated using Laplace and Mohler weights at the initialization layer of the model to enhance the interpretability. Chen et al [24] proposed an interpretable time–frequency network, which embedded a time–frequency convolutional layer in the preprocessing layer to mine the time–frequency features. Dong et al [25] proposed an interpretable multi-scale boosted wavelet network, an interactive channel-attention mechanism was designed to select important fault features. However, the above methods only introduce physical information at the initialization stage, without further enhancing the interpretability at the deep structure of the models. Moreover, although the attention mechanism can enhance important features and suppress redundant information, its interpretability is still limited due to the lack of physical knowledge.

In practical industrial applications, the reliability of intelligent diagnostic models is crucial, their interpretability helps to reveal their internal working mechanisms and increase the transparency of decision-making. However, the interpretability is usually limited to specific layers, and the entire model structure is not designed for the interpretability. In addition, the feature learning performance of the above models is decreased dramatically in noisy environments, affecting the fault diagnosis accuracy. Therefore, it is necessary to develop an interpretable intelligent fault diagnosis method for rotating machinery under noise environments and small sample conditions. Inspired by wavelet transform and Kolmogorov-Arnold Network (KAN), this paper proposes a fault diagnosis method based on wavelet denoising and KANTransformer (WD-KANTF) to improve the robustness and interpretability. The main contributions of this paper are described as follows:

(1) An adaptive wavelet denoising layer is developed to enhance the interpretability of the model. It transforms the input signals from the time domain to the wavelet domain by wavelet transform, which significantly enhances the interpretability. In addition, this layer has a smoothed soft-thresholding and fusion strategy, which can adaptively identify and focus on useful features while effectively suppressing redundant information, thus enhancing the robustness and noise immunity of the model.

(2) To address the limitations of the traditional fully connected layer in nonlinear feature extraction and to improve the interpretability of the deep structure, KANTransformer is designed, which introduces a learnable activation function in the linear layer and significantly enhances the nonlinear feature extraction capability.

(3) An end-to-end intelligent fault diagnosis framework that combines the advantages of CNN and Transformer is proposed to achieve multi-level features learning from local features to global features. Compared with the traditional single CNN or Transformer, this framework has better fault identification performance and stronger robustness under small sample conditions and noise environments.

The rest of the paper is organised as follows. Section 2 introduces the relevant theories. Section 3 describes WD-KANTF in detail. Section 4 performs the experimental validation and analysis. Section 5 concludes this paper.

2. Relevant theories

2.1. Discrete wavelet transform (DWT)

Wavelet transform achieves multi-resolution analysis for vibration signals by dynamically adjusting the size of the time window. DWT decomposes the input signals into high-frequency and low-frequency signals by wavelet function and scale function [1,26]. Meanwhile, the number of signals is multiplied by continuously re-decomposing the low-frequency signals. The DWT decomposition formula is described as follows:

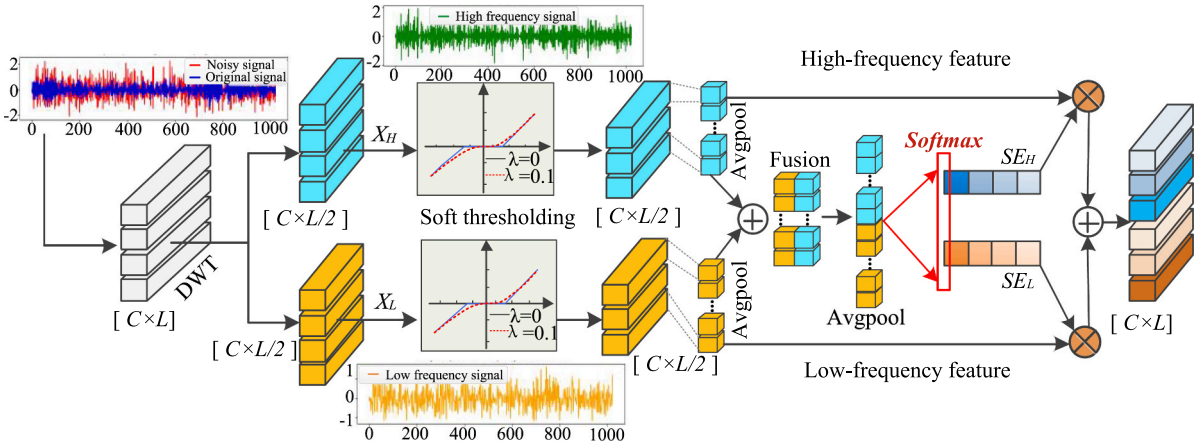


Fig. 1. Structure of adaptive wavelet denoising layer.

$$F_{DWT}(j,k)=\frac{1}{\sqrt{2^j}}x(t)\cdot\psi((t-k2^j)/2^j)\cdot dt \tag{1}$$

where j denotes the scale parameter, which is used to control the scaling of the wavelet function. k denotes the translation parameter, which is used to control the position of the wavelet function on the time axis. $x(t)$ is the input signal. The daubechies (db) wavelet is used in this paper, whose wavelet function and scale function are defined as follows:

$$\varphi_{j,k}^{db}=\sqrt{2}\sum_n L(n)\cdot\varphi(2t-n) \tag{2}$$

$$\psi_{j,k}^{db}=\sum_n H(n)\cdot\varphi(2t-n) \tag{3}$$

where $\varphi_{j,k}^{db}$ denotes the scale function, $\psi_{j,k}^{db}$ denotes the wavelet function, $L(n)$ denotes the low pass filter, $H(n)$ denotes the high pass filter, n denotes the index of the filter.

2.2. Multi-head self-attention mechanism (MHSA)

MHSA is a critical part for the Transformer framework, which can automatically capture the interdependencies in the data and mine the global feature information [27–29]. Furthermore, MHSA consists of multiple parallel scale dot-product attention. Specifically, it generates attention weights by query matrix, key matrix and value matrix. The output is obtained by cascading them to focus on the feature information in different dimensions. Its formulae are described as follows:

$$Att(q,k,v)=\mathrm{Softmax}(\frac{q\cdot k^T}{\sqrt{d_k}})\cdot v \tag{4}$$

$$F_{MHSA}(x)=\mathrm{Linear}(Concate(Att_i)) \tag{5}$$

where q denotes the query matrix, k denotes the key matrix, v denotes the value matrix, T denotes the transpose operation, Softmax$(\cdot)$ denotes the Softmax function, $Concate(\cdot)$ denotes the cascade operation and Linear$(\cdot)$ denotes the linear layer.

2.3. Kolmogorov-Arnold network (KAN)

KAN is based on the Kolmogorov-Arnold theorem, which assumes that any continuous multivariate function can be expressed with the simple univariate function [30,31]. Unlike multi-layer perceptron (MLP), KAN uses the learnable activation function for data processing, which enhances the nonlinear extraction ability of the model and helps the model to be well interpretable. KAN can be described as follows:

$$F_{KAN}(x)=\sum_{q=1}^{2n+1}\varphi_q\left(\sum_{p=1}^{n}\phi_{q,p}(x_p)\right) \tag{6}$$

where φ_q and $\phi_{q,p}$ denote continuous univariate functions, x_p denotes the input signal.

3. Proposed method

In this section, the fault diagnosis framework of WD-KANTF is described in detail. First, the component modules of WD-KANTF are theoretically analysed. Then, the overall structure of WD-KANTF is introduced. Finally, the diagnosis process of WD-KANTF is presented.

3.1. Adaptive wavelet denoising layer

Deep learning has been successfully applied to the field of fault diagnosis and remaining life prediction for rotating machinery due to its powerful representational capability. However, the black-box characteristic of deep learning makes it difficult to understand the working principle. To improve the interpretability of the model, an adaptive wavelet denoising layer (AWDL) is designed, as shown in Fig. 1.

In Fig. 1, DWT is used to extract the low-frequency and high-frequency components from the input signals. Then, the low-frequency and high-frequency components are denoised using smoothed soft thresholding to reduce the interference from the noise environments. Finally, the low-frequency and high-frequency components are adaptively fused using the fusion strategy to retain critical fault features and suppress redundant information. AWDL enhances the feature representation and improves the interpretability through wavelet transform. In addition, the soft thresholding technique is used to improve the robustness of the model. AWDL can be divided into three steps.

Step1: The input features are decomposed by one level of DWT to obtain the low frequency and high frequency components, which are described as follows:

$$x_L^J=\sqrt{2}\sum_n L(n)\cdot x(2t-n) \tag{7}$$

$$x_H^J=\sum_n H(n)\cdot x(2t-n) \tag{8}$$

where $x(t)$ denotes the input signal, $L(n)$ denotes the low pass filter, $H(n)$ denotes the high pass filter and J denotes the number of layers of wavelet decomposition. The time-domain signals are converted into low-frequency and high-frequency components under wavelet domain by DWT, which captures the fault features in different frequency bands.

Step2: The noise in the low-frequency and high-frequency compo-

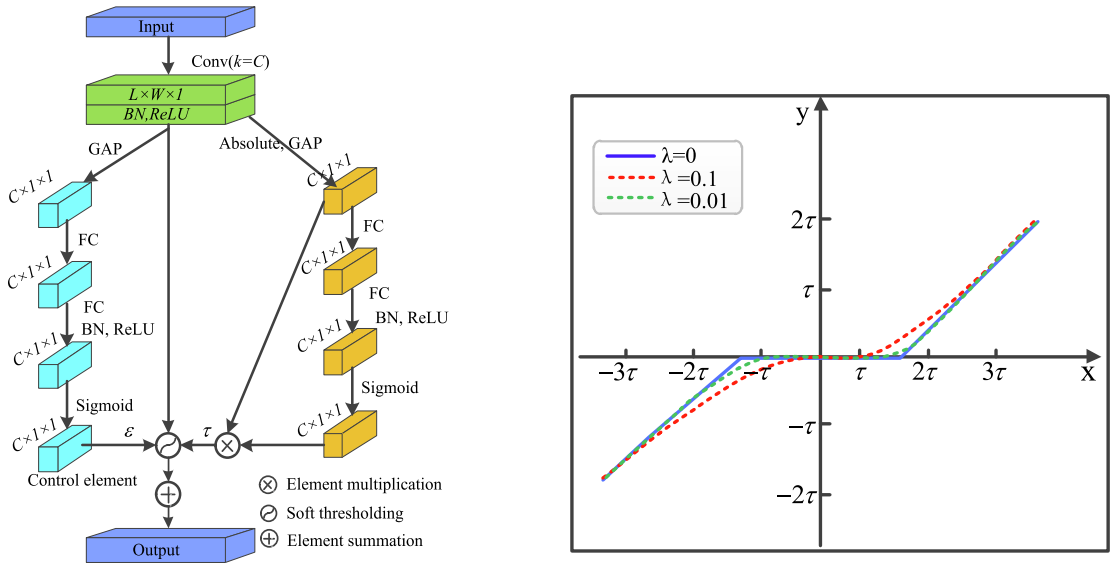


Fig. 2. Smoothed soft threshold function and visualization results.

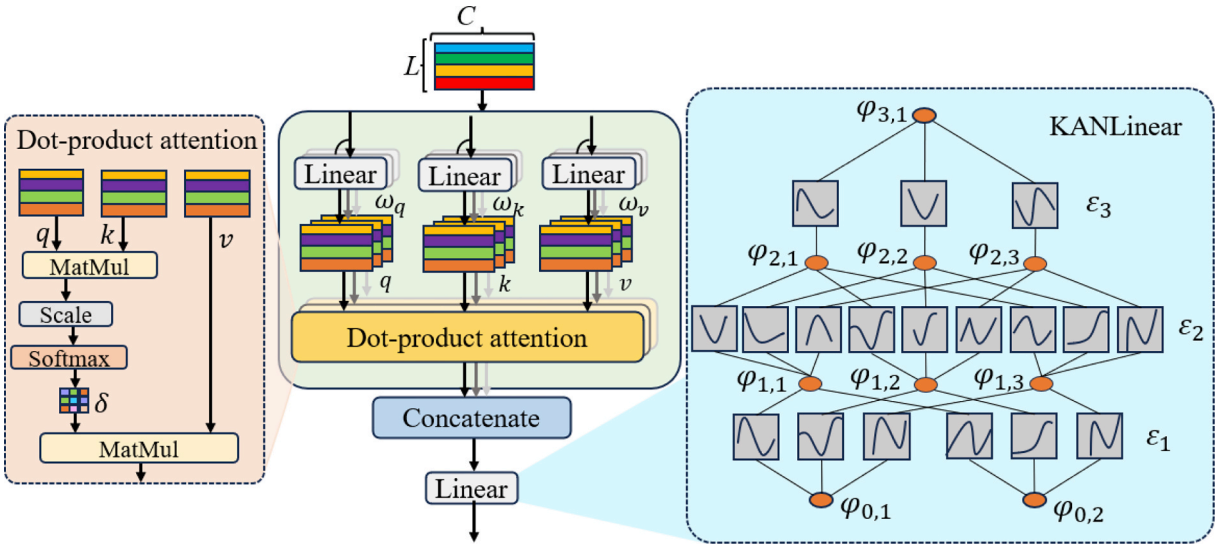


Fig. 3. Schematic structure of the KANT transformer.

nents is filtered using the smoothed soft-threshold function, which improves the noise immunity [32]. Compared with the traditional soft threshold function, the smoothed soft threshold function introduces a control term, which helps the soft threshold function to be transformed from segmented function to curvilinear function, thus preserving the fault features around 0. The structure diagram of the smoothed soft threshold function is shown in Fig. 2. Its formula is described as follows:

$$F_{smooth} = X + \frac{1}{2}\left(\sqrt{(X-\tau)^2+\varepsilon} - \sqrt{(X+\tau)^2+\varepsilon}\right) \tag{9}$$

where X and F_{smooth} are the inputs and outputs of the smoothed soft threshold function, τ is the threshold value and ε is the control term. Thus, the smoothed soft thresholding operation is performed on the high frequency and low frequency components, which can be described as follows:

$$x_L = F_{smooth}(x_L^J) \tag{10}$$

$$x_H = F_{smooth}(x_H^J) \tag{11}$$

where F_{smooth} is the smoothed soft threshold function.

Step3: The low-frequency and high-frequency components are adaptively fused using the fusion strategy. Specifically, an average pooling operation is performed on the low-frequency and high-frequency components, and the features obtained after the average pooling operation are fused. The formula is described as follows:

$$F_{fusion} = f_{avg}(x_L) + f_{avg}(x_H) \tag{12}$$

where $f_{avg}(\cdot)$ is the average pooling and F_{fusion} is the fused output. Then, the average pooling is further performed on F_{fusion} to generate the channel information Z_c, and two fully connected layer mappings are performed on the channel information to obtain two branch features. The important features of different branches are weighted using the Softmax function to obtain the soft attention vectors for low frequency features and high frequency features. The formulas are described as follows:

$$Z_c = f_{avg}(F_{fusion}) \tag{13}$$

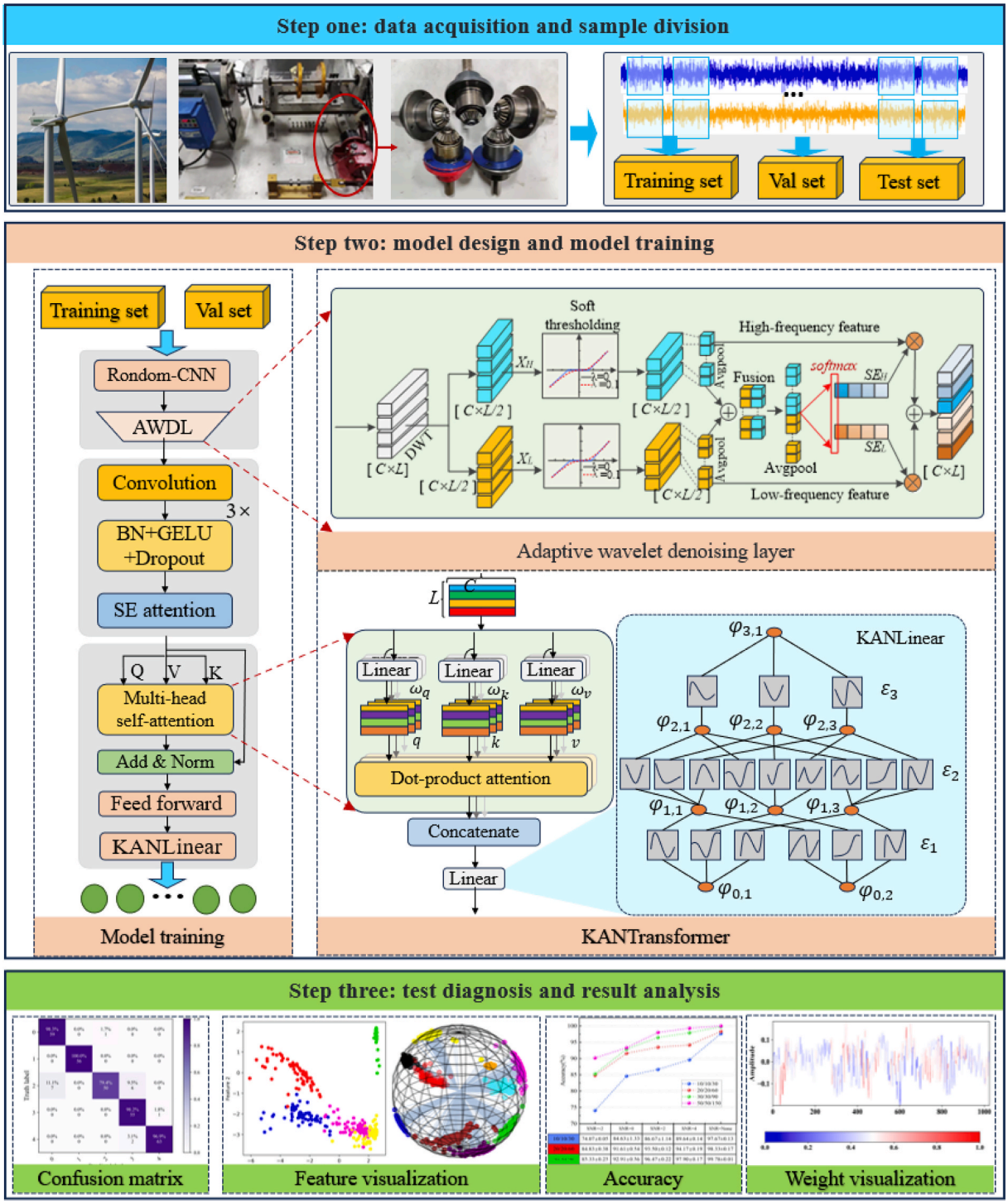


Fig. 4. Fault diagnosis flowchart of the WD-KANTF.

$$F_H = \frac{e^{H_i \cdot Z_c}}{e^{H_i \cdot Z_c} + e^{L_i \cdot Z_c}} \tag{14}$$

$$F_L = \frac{e^{L_i \cdot Z_c}}{e^{H_i \cdot Z_c} + e^{L_i \cdot Z_c}} \tag{15}$$

where $Z_c \in d \times 1$, d is the number of channels in the AWDL. $H, L \in R^{d \times d}$, H_i and L_i are the $i-$th column elements in the matrix. $F_H + F_L = 1$. Finally, the soft attention vectors of low-frequency features and high-frequency features are multiplied with their inputs to complete the recalibration and fusion of the features. The formula is described as follows:

$$X = x_L \otimes F_L + x_H \otimes F_H \tag{16}$$

where $\otimes$ denotes the multiplication operation, x_L and x_H are the inputs of the low and high frequency components, respectively, F_L and F_H are the recalibrated low and high frequency components.

3.2. KANTransformer

In recent years, Transformer has been widely used in the field of

Table 1
Division of case A.

Load(hp)	Bearing state	Fault Diameter	Label	Train /Valid /Test samples
1790 (0hp)	Normal	–	0	10/10/30
1772	Ball	0.007	1	20/20/60
(1hp)		0.014	2	30/30/90
1750		0.021	3	50/50/150
(2hp)	Inner	0.007	4	
1730		0.014	5	
(3hp)		0.021	6	
	Outer-6	0.007	7	
		0.014	8	
		0.021	9	

intelligent fault diagnosis due to its powerful scalability and global dependency capability. MLP of the transformer usually consists of two linear layers and a nonlinear activation function, which leads to the insufficient nonlinear learning ability of the model. Furthermore, each neuron of each layer needs to connect with all the neurons in the previous layer, leading to the dramatic increase of model parameters. This not only increases the computational burden, but also increases the risk of overfitting. Therefore, KANLinear is designed to replace MLP. KANLinear introduces a paradigm shift by placing learnable activation functions on the weights, which not only increases the nonlinear learning ability for the model, but also improves the interpretability. In KANLinear, the learnable activation function consists of multispline functions and basic functions. The specific formulas are described as follows:

$$\varepsilon(x) = \lambda(spline(x) + b(x)) \tag{17}$$

$$spline(x) = \sum_i c_i B_i(x) \tag{18}$$

$$b(x) = \frac{x}{1 + e^{-x}} \tag{19}$$

where λ is a control term, which is used for controlling the overall size of the activation function. B_i is the basic function of the spline. c_i is the control coefficient of the multispline function, which is obtained from the network training. After obtaining a learnable activation function, the output of KANLinear can be described as follows:

$$KANLinear = (\varepsilon_1 \cdot \varepsilon_2 \cdot \varepsilon_3)x \tag{20}$$

where ε_1, ε_2 and ε_3 denote spline functions at different levels, x denotes the input.

Subsequently, KANTransformer is constructed, whose structure is shown in Fig. 3. In Fig. 3, the input features are linearly operated in parallel through the parameter matrices ω_k, ω_q and ω_v to generate the key matrix k, the query matrix q, and the value matrix v. The formulas are described as follows:

$$k = \omega_k x, \quad q = \omega_q x \quad v = \omega_v x \tag{21}$$

where $k,q,v \in R^{N \times C_h}$, x denotes the input features, C denotes the number of channels, N denotes the time index, and h denotes the number of attention mechanisms.

Then, matrix multiplication operation is performed on the query matrix and key matrix, and they are scaled and normalised to obtain the attention weights. The formula is described as follows:

$$att_h = f_{softmax}\left(\frac{q \cdot \kappa^T}{\sqrt{C_h}}\right) \tag{22}$$

where T denotes the transpose matrix and $f_{softmax}(\cdot)$ denotes the Softmax

Table 2
Division of case B.

Load(hp)	Gear state	Label	Train /Valid /Test samples
0hp	Normal bevel gear	0	10/10/30
1hp	Small-end broken fault	1	20/20/60
2hp	Complete tooth fault	2	30/30/90
3hp	Big-end broken fault	3	50/50/150
	Uniform wear	4	

Table 3
Parameters of WD-KANTF.

Name of the layer	operation	Output size
Rondom-CNN	Conv 16@ [15 × 1] , padding = same BN, GELU	[64, 16, 1024]
AWDL	DWT (stage = 1, wavelet = daubechies)	[64, 16, 1024]
First convolutional layer	Conv 32@ [3 × 1] , padding = same BN, GELU, Dropout (0.1)	[64, 32, 1024]
Second convolutional layer	Conv 64@ [3 × 1] , padding = same BN, GELU, Dropout (0.1) Maxpooling (size = 2, stride = 2)	[64, 64, 510]
Third convolutional layer	Conv 128@ [3 × 1] , padding = same BN, GELU, Dropout (0.1) AdaptiveMaxPooling (4)	[64, 128, 510]
KANTransformer	Number of attention heads 2 Number of embedded dimensions 128 Spline order 3, grid size 5,	[64, classes]

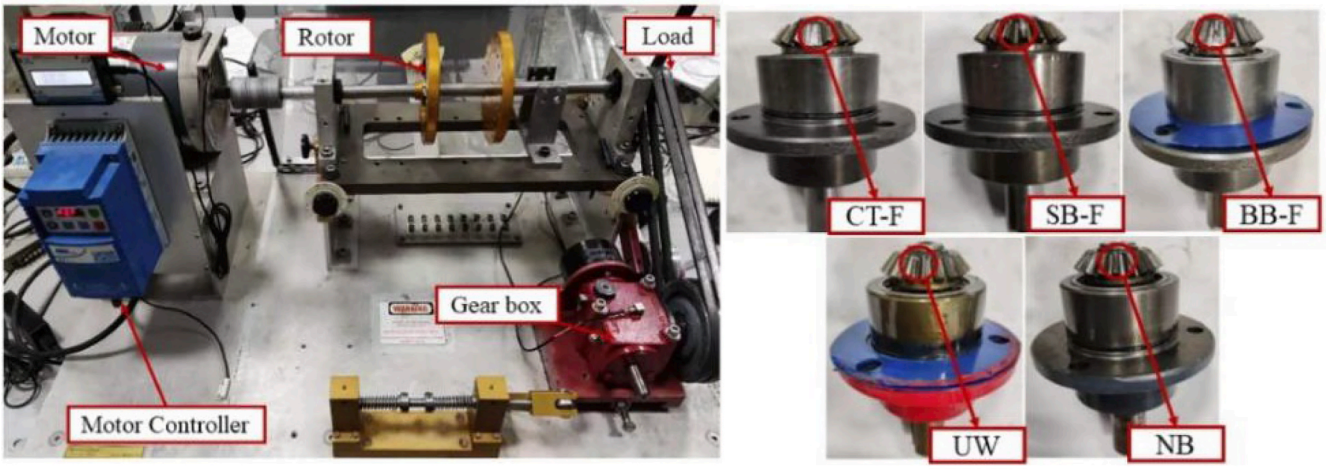
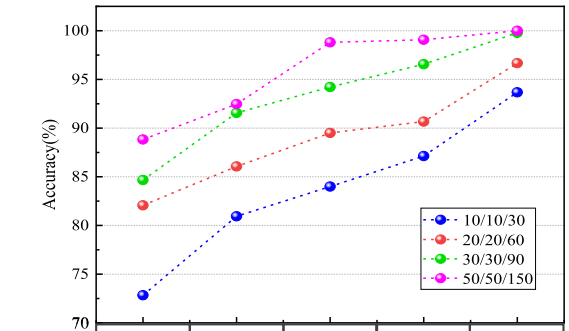
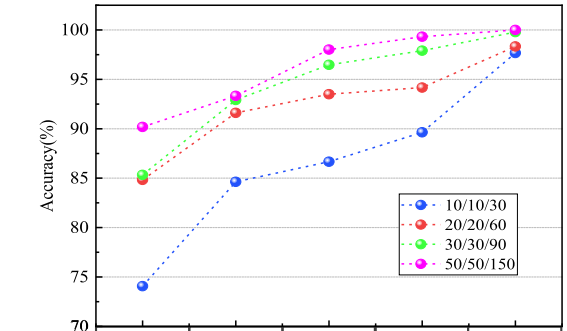


Fig. 5. Test rig for Case B.



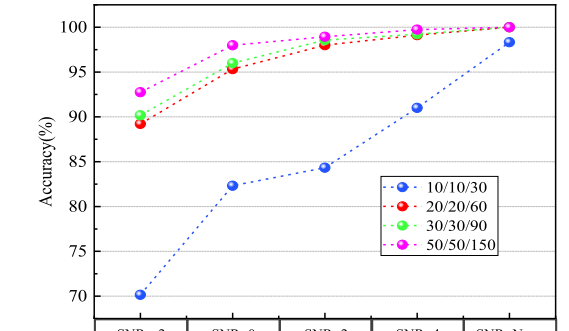
	SNR=-2	SNR=0	SNR=2	SNR=4	SNR=None
10/10/30	72.84±0.23	80.94±1.22	84.00±1.03	87.13±0.35	93.67±0.76
20/20/60	82.07±0.14	86.06±0.64	89.50±0.15	90.67±0.80	96.67±0.45
30/30/90	84.67±0.53	91.56±1.04	94.22±0.23	96.56±0.10	99.78±0.01
50/50/150	88.83±0.76	92.47±0.08	98.80±0.30	99.07±0.36	100.0±0.0

(a) 0HP



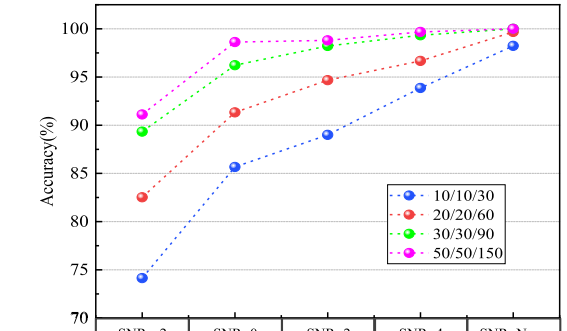
	SNR=-2	SNR=0	SNR=2	SNR=4	SNR=None
10/10/30	74.07±0.05	84.63±1.33	86.67±1.14	89.64±0.14	97.67±0.13
20/20/60	84.83±0.38	91.61±0.54	93.50±0.12	94.17±0.19	98.33±0.17
30/30/90	85.33±0.25	92.91±0.56	96.47±0.22	97.90±0.17	99.78±0.01
50/50/150	90.18±0.43	93.33±0.05	98.01±0.43	99.32±0.05	100.0±0.0

(b) 1HP



	SNR=-2	SNR=0	SNR=2	SNR=4	SNR=None
10/10/30	70.15±1.08	82.32±0.09	84.33±0.15	91.00±0.24	98.33±0.18
20/20/60	89.22±1.45	95.33±0.13	98.00±0.04	99.12±0.20	100.0±0.0
30/30/90	90.17±0.13	96.00±0.16	98.56±0.08	99.22±0.02	100.0±0.0
50/50/150	92.75±0.80	97.99±0.66	98.93±0.08	99.73±0.05	100.0±0.0

(c) 2HP



	SNR=-2	SNR=0	SNR=2	SNR=4	SNR=None
10/10/30	74.14±1.40	85.67±0.07	89.00±0.04	93.86±0.06	98.23±0.13
20/20/60	82.52±0.29	91.33±0.12	94.67±0.31	96.67±0.10	99.67±0.10
30/30/90	89.33±0.21	96.22±0.16	98.22±0.05	99.33±0.18	100.0±0.0
50/50/150	91.11±0.13	98.63±0.30	98.80±0.03	99.67±0.04	100.0±0.0

(d) 3HP

Fig. 6. Fault diagnosis results of WD-KANTF for Case A.

function.

Finally, the attention weights are multiplied with the value matrix and the output of KANTransformer is obtained through KANLinear. The formula is described as follows:

$$X^{output} = Concate(att_h \cdot v) \cdot KANLinear \tag{23}$$

where $Concate(\cdot)$ denotes multiple attention heads for cascading.

3.3. Overall structure

The flowchart of WD-KANTF is shown in Fig. 4. In Fig. 4, WD-KANTF can be summarised into three steps: data acquisition and sample division, model design and model training, fault diagnosis and result analysis.

Step 1: Data acquisition and sample division. Vibration signals are collected from the gearbox and bearing test rigs. The vibration signals are used to divide the data into training set, validation set and test set using sliding window according to 0.2:0.2:0.6.

Step 2: Model design and model training. The input data are passed through convolution and adaptive wavelet denoising layer to obtain local feature information in the wavelet domain. Then, the fully connected layer of CNN is replaced using KANTransformer, which can obtain global feature information. Meanwhile, the nonlinear learning capability of the model is increased by placing learnable activation functions. Finally, the model with the highest accuracy on the valid set is selected as the trained model during the iteration process.

Step 3: Fault diagnosis and result analysis. The test set is input into the trained completed model to obtain the fault diagnosis results. Furthermore, multi-perspective visual analysis is performed using confusion matrix, t-SNE and TGAM.

4. Experimental validation and result analysis

In this section, two datasets are used to evaluate the fault diagnosis effectiveness of WD-KANTF. The computer configuration is AMD Riptide 5-4600H processor and the experimental framework is python-torch.

4.1. Dataset description

4.1.1. Case A

The bearing dataset is derived from the Case Western Reserve University bearing test rig [33,34]. The test rig consists of a motor, a torque transducer, and the tested bearing. The type of the tested bearing is SKF6205 and the sampling frequency is 12 KHz. A total of 0HP, 1HP, 2HP, and 3HP (horsepower) data are collected from the test rig. The fault states are rolling, inner ring and outer ring 6o'clock, and the fault diameters are 0.007, 0.014 and 0.021 in.. Nine classes of fault data and one class of health data are selected. The samples are selected using sliding window and its size is 1024. Table 1 shows the detailed division.

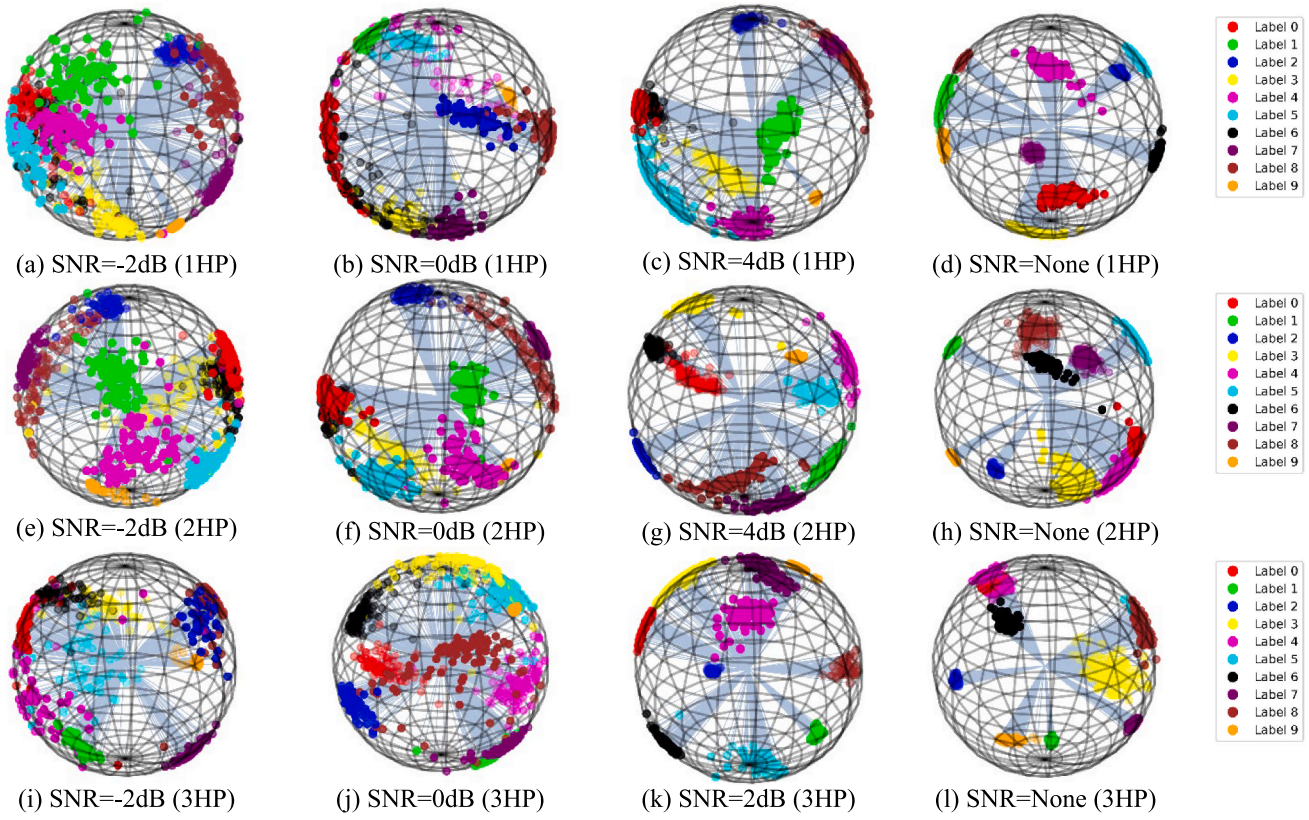


Fig. 7. Visualization results of t-SNE for Case A.

Table 4
Fault diagnosis results of Case B (%).

	0HP				1HP			
	10/10/30	20/20/60	30/30/90	50/50/150	10/10/30	20/20/60	30/30/90	50/50/150
SNR=None	98.33 ± 0.07	99.00 ± 0.05	99.67 ± 0.11	100.0 ± 0.0	98.67 ± 0.11	99.67 ± 0.02	100.0 ± 0.0	100.0 ± 0.0
SNR = −4dB	84.00 ± 0.38	83.67 ± 0.23	87.56 ± 0.32	88.81 ± 0.12	84.42 ± 0.13	86.89 ± 0.22	88.52 ± 0.10	89.78 ± 0.24
SNR = −2dB	92.15 ± 0.66	94.00 ± 0.09	93.48 ± 0.26	94.07 ± 0.06	89.33 ± 0.05	92.65 ± 0.30	93.60 ± 0.22	95.65 ± 0.16
SNR = 0 dB	94.67 ± 0.08	96.07 ± 0.14	97.88 ± 0.04	98.40 ± 0.02	90.50 ± 0.10	97.67 ± 0.18	98.84 ± 0.36	98.68 ± 0.24
SNR = 2 dB	95.29 ± 0.27	97.67 ± 0.03	98.95 ± 0.14	99.27 ± 0.03	92.36 ± 0.23	98.41 ± 0.02	99.07 ± 0.12	99.33 ± 0.04
SNR = 4 dB	97.38 ± 0.17	98.93 ± 0.18	99.36 ± 0.04	99.47 ± 0.24	96.43 ± 0.08	99.00 ± 0.05	99.85 ± 0.02	99.91 ± 0.05
	2HP				3HP			
	10/10/30	20/20/60	30/30/90	50/50/150	10/10/30	20/20/60	30/30/90	50/50/150
SNR=None	98.00 ± 0.07	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	98.25 ± 0.04	99.77 ± 0.12	100.0 ± 0.0	100.0 ± 0.0
SNR = −4dB	86.00 ± 0.19	89.67 ± 0.08	91.06 ± 0.22	93.08 ± 0.22	81.55 ± 0.24	89.00 ± 0.18	88.78 ± 0.70	90.96 ± 0.33
SNR = −2dB	93.33 ± 0.02	94.33 ± 0.11	95.56 ± 0.45	95.87 ± 0.03	85.03 ± 0.74	93.96 ± 0.04	95.03 ± 0.15	96.93 ± 0.05
SNR = 0 dB	94.00 ± 0.32	96.67 ± 0.03	96.81 ± 0.17	98.53 ± 0.04	92.00 ± 0.20	95.64 ± 0.30	96.56 ± 0.07	98.51 ± 0.16
SNR = 2 dB	95.67 ± 0.34	98.00 ± 0.01	97.11 ± 0.06	99.73 ± 0.12	95.08 ± 0.33	98.48 ± 0.17	98.89 ± 0.22	99.67 ± 0.05
SNR = 4 dB	97.14 ± 0.56	98.98 ± 0.03	99.06 ± 0.02	100.0 ± 0.0	98.00 ± 0.12	99.06 ± 0.07	99.38 ± 0.46	100.0 ± 0.0

4.1.2. Case B

The gear dataset is collected by ourselves, and the test platform is shown in Fig. 5. The platform mainly consists of an asynchronous motor, a loading device, a gearbox and an inverter controller. The sensor is the shear acceleration sensor and the sampling frequency is 25.6 KHz. The test object is the bevel gears, and the loading device is adjusted to obtain the data of 0HP, 1HP, 2HP and 3HP at 30 Hz. The fault states are complete tooth fault (CT-F), small end breakage fault (SB-F), big end breakage fault (BB-F), uniform wear (UW) and normal bevel gear state (NB). Table 2 shows the detailed division.

4.1.3. Parameters of the model and experimental setup

Table 3 shows the parameters of WD-KANTF, these parameters are mainly determined by cross-validation. In addition, the network parameters are updated using the Adam algorithm. The learning rate is set to 0.0005 and the batch size is 64.

4.2. Diagnosis results

4.2.1. Diagnosis results and analysis of Case A

In practical industrial scenarios, rotating machinery is usually subjected to noise environments. Noise environments make the fault features more complex. Therefore, fault diagnosis under noise environments and small sample conditions is a challenging task. The small sample conditions are set as 50, 100, 150 and 250 samples. The noise environments include signal-to-noise ratio (SNR) of −2dB, 0 dB, 2 dB and 4 dB. SNR=None indicates that noise is not added. Fig. 6 shows the experimental results.

In Fig. 6, the diagnostic accuracy of WD-KANTF has an obvious increasing trend by increasing SNR from −2dB to 4 dB. Specifically,

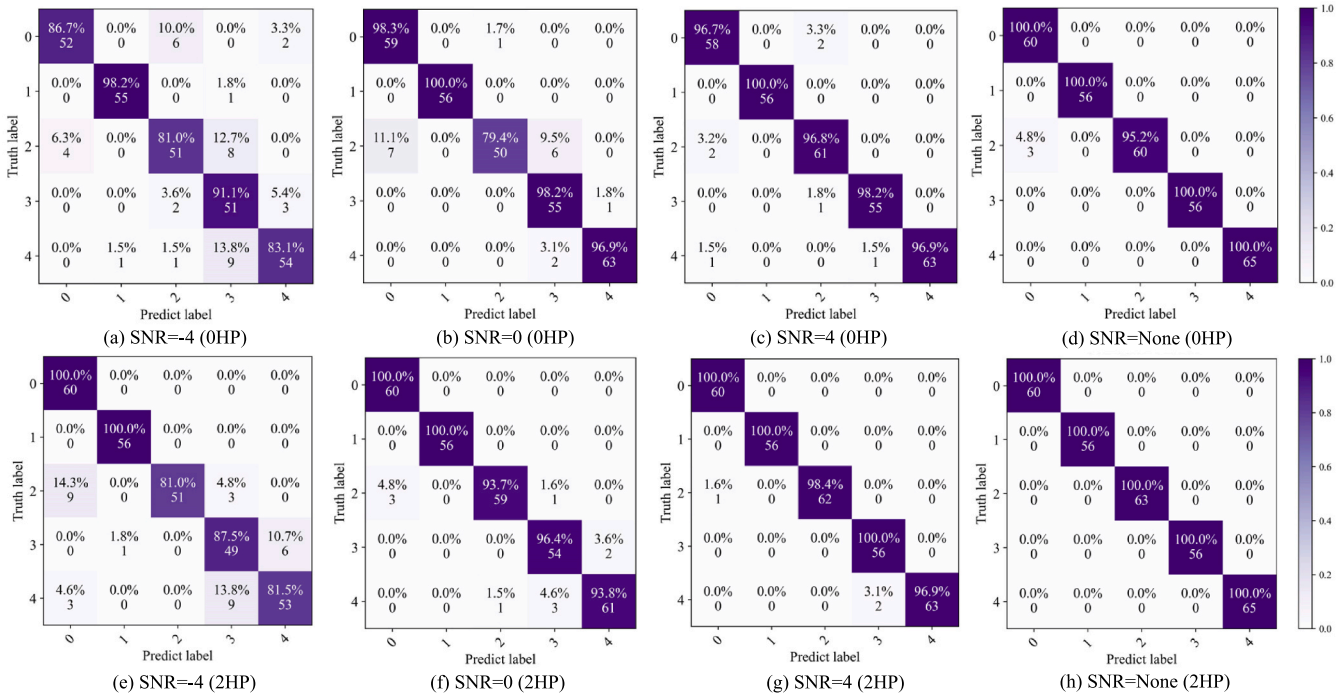


Fig. 8. Confusion matrix results for Case B.

when noise is not added, the diagnostic accuracy of WD-KANTF is above 93 % in four small samples conditions. The diagnostic accuracy of WD-KANTF is 100 % at the sample size of 250. In Fig. 6 (a), the fault diagnosis accuracy of WD-KANTF is 72.84 % when the number of samples is 50 and SNR = −2dB. The diagnostic accuracy of WD-KANTF is 87.13 % as SNR = 4 dB is increased, which shows that WD-KANTF can mine fault features. This is because the proposed method uses KANTransformer in the fully connected layer, which can assign different weights to the deep features, thus enhancing the useful features and suppressing the redundant information.

In order to explore the diagnostic performance, the last layer of WD-KANTF is visualised and analysed by using 1HP and 2HP, the results are shown in Fig. 7. The features extracted by the model are uniformly scaled to the unit sphere. WD-KANTF exhibits low feature clustering on the sphere surface at SNR = −2dB, which indicates that the strong noise environments make the fault features more complex, resulting in insufficient feature extraction capability. However, with the increasing of SNR, the interference in the noise environments for fault features is gradually weakening, and WD-KANTF can extract fault features more easily. For example, Fig. 7(b) and Fig. 7(c) show better feature clustering. When noise is not added, WD-KANTF can cluster different fault features in different regions and the fault features are clearly identifiable. For example, Fig. 7(d), (h) and (l) show the best clustering results, which highlight the excellent fault diagnosis performance for WD-KANTF under small sample conditions.

4.2.2. Diagnosis results and analysis of Case B

In order to verify the generalization performance of WD-KANTF, the gear dataset is used for validation. The noise environments include −4dB, −2dB, 0 dB, 2 dB, and 4 dB. The small sample conditions include 50, 100, 150, and 250 samples. Table 4 shows the fault diagnosis results. The following conclusions can be drawn from Table 4: (1) WD-KANTF has superior feature extraction capability under noise environments and small sample conditions. Specifically, the fault diagnosis accuracy of WD-KANTF is above 84 % at SNR = −4dB and 10/10/30 task. This indicates that the adaptive wavelet denoising layer can effectively improve the anti-noise performance of the model. (2) The diagnostic accuracy of WD-KANTF increases with the increasing of SNR, and its diagnostic accuracy exceeds 97 % at SNR = 4 dB. (3) When the number of samples is 250, the diagnostic accuracy of WD-KANTF is 100 %. When the number of samples is reduced from 250 to 50, its diagnostic accuracy is still above 98 %.

In order to observe the classification of each fault for WD-KANTF, the confusion matrix is used for visualization and analysis. The results are shown in Fig. 8, the horizontal and vertical axes represent the predicted labels and true labels, respectively. In Fig. 8(a), all the five fault classes are misclassified, and lots of fault samples of label 2 and label 4 are incorrectly identified. This is due to the fact that the vibration signals of the gears are affected by strong noise, which make the fault features more complex. In Fig. 8(b)-Fig. 8(d), the misclassification rates of the five fault categories are decreased with the increasing of SNR. WD-KANTF misclassifies the three labels 2 as label 0 at SNR=None, while all other fault classes are accurately classified. Furthermore, Fig. 8(e)-Fig. 8(h) show the fault diagnosis results at 2HP. In Fig. 8(e), three fault classes for labels 2, 3 and 4 are misclassified. However, the diagnostic accuracies of WD-KANTF for the five fault classes improve with the increasing of SNR. The five fault classes are correctly classified and its fault diagnosis accuracy is 100 % at SNR=None.

4.3. Fault diagnosis results and analysis of comparison methods

In order to completely evaluate the diagnostic performance of WD-KANTF, eight state-of-the-art methods are selected for comparison, which include EWSNet [35], TFN [24], DLWCB [36], LiConvFormer [37], MRSCNN [38], AMARSN [32], DRSwin-ST [39], and CKG [40]. EWSNet, TFN and DLWCB utilise wavelet prior knowledge for convolutional weight initialization. LiConvFormer is a lightweight fault diagnosis model. MRSCNN uses soft thresholding for noise immunity research. AMARSN is a multi-scale residual shrinkage method with global and local denoising. DRSwin-ST is an adaptive thresholding denoising model based on Swin Transformer. CKG is a classification-based knowledge-guided few-shot fault diagnosis model, which utilises an incomplete multi-kernel clustering algorithm for few-shot fault diagnosis. The above comparison methods only adjust the input sample size and have the same structure as mentioned in the published papers. All the methods have been subjected for five experiments to ensure the

Table 5
Fault diagnosis results of Case A under small sample conditions.

	0HP			1HP		
	10/10/30	20/20/60	30/30/90	10/10/30	20/20/60	30/30/90
EWSNet	66.67 ± 0.12	86.78 ± 1.24	93.75 ± 0.55	70.00 ± 0.22	88.36 ± 0.14	94.43 ± 0.36
TFN	78.12 ± 0.04	86.86 ± 0.33	92.19 ± 0.55	79.16 ± 1.06	94.64 ± 0.77	95.75 ± 0.14
LiConvFormer	75.00 ± 0.64	88.36 ± 0.76	92.81 ± 0.35	76.56 ± 1.06	90.79 ± 0.12	94.12 ± 0.66
DLWCB	75.78 ± 0.56	89.32 ± 1.04	95.44 ± 0.66	78.12 ± 0.21	89.05 ± 0.75	93.04 ± 0.28
MRSCNN	68.33 ± 1.14	78.00 ± 0.44	83.33 ± 0.61	79.83 ± 0.17	86.83 ± 0.11	92.02 ± 0.88
AMARSN	65.67 ± 0.32	74.35 ± 0.77	80.22 ± 0.45	69.67 ± 0.96	79.17 ± 0.06	86.22 ± 1.42
DRSwin-ST	74.22 ± 0.99	84.93 ± 0.42	95.64 ± 0.22	77.52 ± 0.21	84.09 ± 1.23	92.77 ± 0.36
CKG	86.73 ± 0.22	90.09 ± 0.17	97.31 ± 0.93	92.41 ± 0.52	96.60 ± 0.64	96.74 ± 0.35
WD-KANTF	93.76 ± 0.76	96.67 ± 0.45	99.78 ± 0.01	97.67 ± 0.13	98.33 ± 0.17	99.78 ± 0.01
	2HP			3HP		
	10/10/30	20/20/60	30/30/90	10/10/30	20/20/60	30/30/90
EWSNet	77.89 ± 0.33	84.64 ± 0.21	89.66 ± 1.02	68.67 ± 0.17	85.56 ± 0.44	87.67 ± 0.94
TFN	78.66 ± 0.45	86.39 ± 0.77	93.75 ± 0.22	75.39 ± 0.99	86.13 ± 0.43	94.52 ± 0.87
LiConvFormer	77.34 ± 0.32	81.45 ± 0.66	94.80 ± 0.56	69.71 ± 1.07	77.53 ± 0.43	86.92 ± 0.88
DLWCB	79.02 ± 0.21	84.03 ± 0.95	95.60 ± 0.49	77.66 ± 0.12	87.12 ± 0.23	93.75 ± 0.16
MRSCNN	80.95 ± 0.44	83.86 ± 0.74	90.87 ± 0.12	77.67 ± 0.33	82.32 ± 0.21	88.62 ± 0.75
AMARSN	70.33 ± 0.74	78.14 ± 0.33	88.67 ± 0.55	71.00 ± 0.26	78.33 ± 0.67	85.75 ± 0.04
DRSwin-ST	74.21 ± 0.05	83.72 ± 0.18	86.05 ± 0.10	73.88 ± 0.68	79.41 ± 0.54	86.49 ± 0.14
CKG	90.68 ± 0.10	91.14 ± 0.97	96.09 ± 0.33	91.87 ± 0.32	96.96 ± 0.24	94.81 ± 0.77
WD-KANTF	98.33 ± 0.18	100.0 ± 0.0	100.0 ± 0.0	98.23 ± 0.13	99.67 ± 0.10	100.0 ± 0.0

Table 6
Fault diagnosis results of Case A in noise environments.

	SNR = −4dB			SNR = −2dB		
	10/10/30	20/20/60	30/30/90	10/10/30	20/20/60	30/30/90
EWSNet	44.67 ± 0.74	46.50 ± 0.42	50.22 ± 0.13	47.68 ± 0.10	55.33 ± 0.24	59.89 ± 0.55
TFN	66.33 ± 0.42	72.66 ± 0.12	75.56 ± 0.66	69.17 ± 1.02	72.15 ± 0.96	75.58 ± 1.24
LiConvFormer	57.98 ± 0.18	61.00 ± 1.07	66.75 ± 0.35	57.00 ± 1.22	65.17 ± 0.45	68.67 ± 0.72
DLWCB	58.44 ± 0.36	56.32 ± 0.11	61.78 ± 0.12	60.02 ± 0.44	70.07 ± 0.65	78.78 ± 0.18
MRSCNN	56.12 ± 0.08	62.44 ± 0.14	67.05 ± 0.46	60.44 ± 0.31	67.50 ± 0.66	70.27 ± 0.11
AMARSN	52.67 ± 0.92	56.56 ± 0.95	65.22 ± 0.27	55.50 ± 0.13	60.00 ± 0.84	67.80 ± 0.92
DRSwin-ST	59.31 ± 0.13	64.15 ± 0.57	72.07 ± 0.67	62.80 ± 0.45	67.39 ± 0.98	75.05 ± 0.77
CKG	61.13 ± 0.27	74.09 ± 0.13	80.33 ± 0.33	62.41 ± 0.02	74.60 ± 0.04	79.74 ± 0.76
WD-KANTF	76.67 ± 0.59	80.31 ± 1.22	86.32 ± 0.74	74.07 ± 0.05	84.83 ± 0.38	90.18 ± 0.43
	SNR = 0 dB			SNR = 2 dB		
	10/10/30	20/20/60	30/30/90	10/10/30	20/20/60	30/30/90
EWSNet	60.00 ± 0.34	75.67 ± 1.04	80.78 ± 0.23	64.68 ± 0.64	78.40 ± 0.22	83.38 ± 0.33
TFN	68.67 ± 0.02	79.67 ± 0.20	85.11 ± 0.21	71.43 ± 0.21	82.09 ± 0.32	88.88 ± 0.53
LiConvFormer	62.00 ± 1.77	77.50 ± 0.13	83.56 ± 0.37	65.62 ± 0.11	77.83 ± 0.32	84.92 ± 0.02
DLWCB	65.96 ± 0.11	72.50 ± 0.19	84.40 ± 1.18	68.59 ± 0.25	78.62 ± 0.84	88.75 ± 0.34
MRSCNN	59.67 ± 0.88	72.88 ± 1.76	85.35 ± 0.24	63.00 ± 0.12	74.03 ± 0.16	86.38 ± 0.23
AMARSN	63.33 ± 0.44	66.83 ± 0.76	70.22 ± 0.47	64.11 ± 0.52	70.11 ± 0.55	82.33 ± 0.47
DRSwin-ST	68.70 ± 0.16	78.13 ± 0.94	76.10 ± 0.88	74.65 ± 0.64	81.23 ± 0.79	84.45 ± 0.58
CKG	70.68 ± 0.18	80.14 ± 0.17	85.09 ± 0.43	75.07 ± 0.37	84.06 ± 0.24	89.01 ± 0.07
WD-KANTF	84.63 ± 1.33	91.61 ± 0.54	92.91 ± 0.56	86.67 ± 1.14	93.50 ± 0.12	96.47 ± 0.22

stability of the methods.

4.3.1. Comparative experimental results for Case A

Samples of 50, 100 and 150 are selected for fault diagnosis research, the experimental results are shown in Table 5. In Table 5, the average diagnosis accuracy of WD-KANTF is 98.51 % at 0HP, 1HP, 2HP and 3HP. The average diagnostic accuracy of WD-KANTF is improved by 15.67 % and 11.72 % compared to EWSNet and TFN, which indicates that the introducing physical knowledge in the model helps to capture critical features. However, the insufficient number of training samples resulted in degradation of fault diagnosis performance for EWSNet and TFN. Compared with LiConvFormer, the average diagnostic accuracy of WD-KANTF is increased by 14.73 %. This is due to the fact that LiConvFormer is a lightweight model and its feature extraction capability is insufficient under small sample conditions. Compared with DLWCB, MRSCNN, AMARSN and DRSwin-ST, the average diagnostic accuracy of WD-KANTF is increased by 12.02 %, 15.80 %, 21.22 % and 15.77 %, respectively, which indicates that the KANTransformer can capture global fault features and improve fault identification accuracy of WD-KANTF under small sample conditions. Compared with CKG, the average diagnostic accuracy of WD-KANTF is improved by 5.06 %, which indicates that the few-shot learning strategy can improve the fault diagnosis performance, but the average diagnosis accuracy of CKG is still lower than the proposed method due to the lack of physical knowledge guidance and overfitting problem.

To further verify the diagnostic performance of WD-KANTF in noise environments, sample conditions of 20/20/60 at 1 HP are selected for experimental verification, and the results are shown in Table 6. In Table 6, with the increasing of SNR, the fault diagnosis accuracies of all methods are gradually improved, which indicate that the interference in the noise environments is decreasing. Under different noise environments, the diagnostic accuracy of WD-KANTF is better than that of other compared methods. Especially, WD-KANTF still maintains the high diagnostic accuracy under strong noise environments. For example, the diagnostic accuracy of WD-KANTF is 76.67 % under SNR = −4dB and 10/10/30 task, which indicates that WD-KANTF has strong noise immunity. In addition, the diagnostic performance of CKG and TFN is the second best. For example, the diagnostic accuracy of CKG is higher than that of TFN at SNR = 0 dB and SNR = 2 dB. TFN has higher diagnostic accuracy at SNR = −4dB and −2dB. MRSCNN and AMARSN have the worst diagnostic results under strong noise environments, which indicates that small sample conditions have large influence on the diagnostic performance of the model. In conclusion, the fault diagnosis accuracy of WD-KANTF in different noise environments is significantly higher than that of other methods, exhibits superior noise immunity under strong noise environments and small sample conditions.

In order to visualise the feature extraction capability of different methods, SNR = 2 dB and 20/20/60 are selected for t-SNE visualization. Fig. 9 shows the t-SNE visualization results for different methods. In Fig. 9, the faults of different classes are fused together without obvious boundaries for EWSNet and LiConvFormer, which indicates that the feature extraction ability of EWSNet and LiConvFormer is insufficient in noise environments, resulting in poor diagnostic accuracy for them. The boundaries of most of the different labels can be separated effectively for

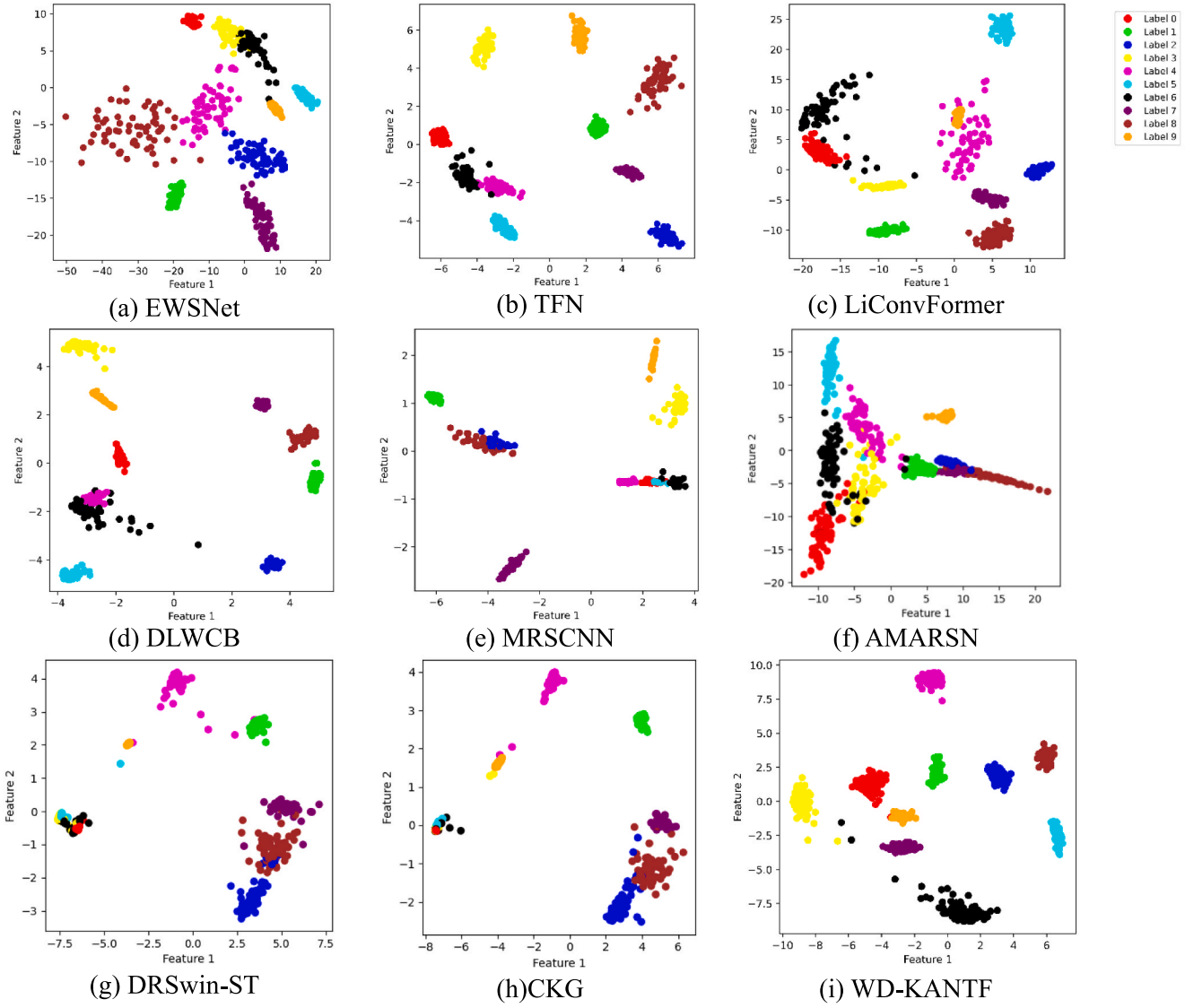


Fig. 9. Visualisation results of t-SNE for different methods in Case A.

TFN, DLWCB and MRSCNN, but label 4 and label 6 are mixed and cannot be separated. For AMARSN, there is no classification boundary for different classes and they are mixed, which indicates that AMARSN has poor feature extraction capability and cannot identify different fault classes. For DRSwin-ST and CKG, most of the classes can be separated effectively, but it is difficult to cluster for labels 0, 5 and 6. This is due to the similarity of fault classes for labels 5 and 6, which is difficult to be recognised. Ten labels can be completely separated for WD-KANTF, which indicates that the same labels are clustered together, and the different label classes are effectively separated. This shows that WD-KANTF has superior feature extraction capability and good robustness under small sample conditions and noise environments.

4.3.2. *Comparative experimental results for Case B*

In order to verify the generalization performance of WD-KANTF, the gearbox dataset is selected for experimental validation and analysis. The experimental results are shown in Table 7. In Table 7, the diagnostic accuracies of all methods show overall increasing trend as the sample numbers increase from 50 to 150, which indicates that the number of training samples has the critical impact on the feature extraction ability, and the diagnostic accuracy of the methods is improved with the increasing number of training samples. In addition, the diagnostic accuracy of WD-KANTF is above 98 % under different small sample conditions. Especially, the diagnostic accuracy is 100.0 % at 2HP and 20/20/60 task, which shows that WD-KANTF has strong fault diagnosis capability. The diagnostic accuracies of CKG and TFN decrease significantly at 10/10/30 task, which indicates that their performance and stability of fault diagnosis are poor. LiConvFormer, MRSCCNN, DLWCB and DRSwin-ST have lower diagnostic accuracies than that of WD-KANTF. The diagnostic accuracy of LiConvFormer is 71.88 $\pm$ 0.29 % at 0HP and 10/10/30 task. The above analysis can be concluded that the fault diagnosis ability of WD-KANTF is significantly better than other methods under small sample conditions.

Table 8 is the fault diagnosis results of the model for Case B at 1HP and 20/20/60 task. In Table 8, the fault diagnosis accuracies of both WD-KANTF and comparison methods improves with the increasing SNR. This indicates that noise has the large impact on the fault diagnosis performance, and the strong noise environments would lead to the decrease of the diagnosis performance of the models. In addition, the diagnostic accuracy of WD-KANTF is much higher than that of the comparison methods in different noise environments. For example, the diagnostic accuracy of WD-KANTF is 76.67 % at SNR $= -4$dB and 10/10/30 task. Compared with EWSNet, TFN, LiConvFormer, MRSCCNN, DLWCB, DRSwin-ST, AMARSN, and CKG, the diagnostic accuracy of

Table 7
Fault diagnosis results of Case B under small sample conditions.

	0HP			1HP		
	10/10/30	20/20/60	30/30/90	10/10/30	20/20/60	30/30/90
EWSNet	77.60 ± 0.52	85.12 ± 1.14	90.42 ± 0.66	76.00 ± 1.03	90.01 ± 0.77	95.76 ± 0.14
TFN	78.12 ± 0.16	88.41 ± 1.02	92.90 ± 0.44	78.12 ± 0.21	92.44 ± 0.10	97.04 ± 0.84
LiConvFormer	71.88 ± 0.29	89.10 ± 0.77	93.02 ± 0.12	80.67 ± 1.07	94.33 ± 0.22	97.65 ± 0.16
DLWCB	71.33 ± 1.02	86.59 ± 0.66	89.37 ± 0.32	79.52 ± 0.88	89.32 ± 0.14	94.25 ± 0.61
MRSCNN	76.00 ± 1.24	82.77 ± 0.04	90.55 ± 0.21	76.04 ± 0.12	88.54 ± 0.10	94.64 ± 0.08
AMARSN	76.65 ± 0.77	85.39 ± 0.90	92.05 ± 0.08	77.56 ± 0.75	86.24 ± 0.87	93.43 ± 0.66
DRSwin-ST	75.00 ± 0.98	86.03 ± 0.47	93.09 ± 0.88	74.53 ± 0.06	84.92 ± 0.75	90.34 ± 0.76
CKG	80.66 ± 0.18	89.57 ± 0.08	94.33 ± 0.34	81.09 ± 0.17	88.99 ± 0.55	97.45 ± 0.12
WD-KANTF	98.33 ± 0.07	99.00 ± 0.05	99.67 ± 0.11	98.67 ± 0.11	99.67 ± 0.02	100.0 ± 0.0
	2HP			3HP		
EWSNet	76.67 ± 0.52	88.49 ± 0.96	91.45 ± 0.36	80.67 ± 0.53	83.12 ± 0.26	89.45 ± 0.36
TFN	73.44 ± 0.52	90.31 ± 0.08	91.48 ± 0.18	78.34 ± 0.23	86.26 ± 1.03	93.45 ± 0.13
LiConvFormer	80.00 ± 1.08	88.28 ± 0.54	94.68 ± 0.06	75.50 ± 0.55	92.06 ± 0.10	93.68 ± 0.24
DLWCB	55.09 ± 1.34	86.54 ± 0.22	92.41 ± 0.55	57.47 ± 0.33	77.83 ± 0.78	80.87 ± 0.22
MRSCNN	75.43 ± 0.77	85.09 ± 0.54	91.94 ± 0.33	76.00 ± 0.48	84.88 ± 0.25	89.97 ± 0.76
AMARSN	73.78 ± 0.73	81.30 ± 0.31	87.99 ± 0.81	78.83 ± 0.44	83.08 ± 0.58	87.96 ± 0.09
DRSwin-ST	76.56 ± 0.73	84.38 ± 0.92	90.45 ± 0.25	78.58 ± 0.19	87.09 ± 0.58	90.66 ± 0.76
CKG	84.77 ± 0.42	89.98 ± 0.64	95.33 ± 0.90	81.91 ± 0.17	89.02 ± 0.21	94.42 ± 0.31
WD-KANTF	98.00 ± 0.07	100.0 ± 0.0	100.0 ± 0.0	98.25 ± 0.04	99.77 ± 0.12	100.0 ± 0.0

WD-KANTF is improved by 18.12 %, 10.96 %, 21.34 %, 22.67 %, 21.67 %, 16.76 %, and 12.07 %, respectively. This indicates that WD-KANTF has strong feature extraction capability and can effectively mine fault information. However, the performance of EWSNet, TFN, LiConvFormer, and MRSCCNN is significantly degraded under small sample conditions. Fig. 10 is the t-SNE visualization results at SNR = 2 dB and 20/20/60 task. In Fig. 10, the different fault classes of EWSNet, TFN, LiConvFormer, DLWCB and AMARSN are mixed, and there is no clear separation boundary for each fault class. For DRSwin-ST and CKG, there are clear classification boundaries for the different fault classes, but the distance between the same fault classes is large. For WD-KANTF, the fault classes have obvious separation boundaries and the same faults are clustered together, which means that WD-KANTF has good feature extraction ability under small sample conditions and noise environments, and can effectively mine fault information.

4.3.3. Timeliness and the number of parameters of WD-KANTF

Computational resources and timeliness are crucial for deep learning models. In this section, the computational resources and timeliness of WD-KANTF are investigated. The results are shown in Table 9. In Table 9, it is observed that LiConvFormer has the shortest training time, which is only 84.67 s. DRSwin-ST has the longest training time, which is 594.75 s. LiConvFormer has the lowest number of parameters for 3.22×10^5 and DRSwin-ST has the highest number of parameters for 5.24×10^7. This is due to the fact that LiConvFormer is a lightweight model, which has the smallest number of parameters and the shortest training time among all the methods. DRSwin-ST is designed based on Swin-Transformer, which has the largest number of parameters and the longest training time. In addition, EWSNet, TFN, MRSCNN, AMARSN, and CKG exhibit longer training time and higher number of parameters. However, compared with the other methods, the training time of WD-KANTF is 101.25 s and the number of parameters is 6.41×10^5, which indicates that WD-KANTF spends less time and computational resources to achieve better diagnostic results under small sample conditions and noise environments.

Table 8
Fault diagnosis results of Case B in noise environments.

	SNR = −4dB			SNR = −2dB		
	10/10/30	20/20/60	30/30/90	10/10/30	20/20/60	30/30/90
EWSNet	58.55 ± 1.38	66.77 ± 0.35	75.78 ± 1.57	62.00 ± 1.10	68.67 ± 1.33	79.56 ± 0.31
TFN	65.71 ± 1.02	77.97 ± 0.28	82.22 ± 0.13	72.22 ± 0.32	76.48 ± 0.77	84.53 ± 1.01
LiConvFormer	55.33 ± 1.13	63.33 ± 0.08	64.56 ± 0.17	62.17 ± 0.67	66.77 ± 0.30	71.67 ± 0.11
DLWCB	54.00 ± 1.29	69.67 ± 0.75	75.96 ± 0.88	52.00 ± 0.60	74.33 ± 0.14	77.65 ± 1.02
MRSCNN	50.00 ± 0.73	70.20 ± 0.66	75.56 ± 0.04	67.33 ± 0.46	81.75 ± 1.48	84.67 ± 0.44
AMARSN	55.00 ± 0.05	62.96 ± 0.18	68.05 ± 0.56	60.43 ± 0.90	71.25 ± 0.10	74.55 ± 0.59
DRSwin-ST	59.91 ± 0.65	67.66 ± 0.05	70.34 ± 0.88	62.22 ± 0.64	69.98 ± 0.85	73.39 ± 0.24
CKG	64.60 ± 0.51	68.37 ± 0.57	76.09 ± 0.77	66.09 ± 0.09	73.01 ± 0.99	77.45 ± 0.14
WD-KANTF	84.42 ± 0.13	86.89 ± 0.22	88.52 ± 0.10	89.33 ± 0.05	92.65 ± 0.30	93.60 ± 0.22
	SNR = 0 dB			SNR = 2 dB		
	10/10/30	20/20/60	30/30/90	10/10/30	20/20/60	30/30/90
EWSNet	64.67 ± 0.17	78.33 ± 1.21	86.25 ± 0.34	67.55 ± 0.44	72.61 ± 0.14	84.89 ± 1.50
TFN	75.15 ± 0.92	82.46 ± 0.12	88.98 ± 0.02	77.62 ± 0.32	85.34 ± 0.23	90.04 ± 0.05
LiConvFormer	71.33 ± 0.90	79.45 ± 0.55	87.36 ± 0.44	69.57 ± 0.42	79.67 ± 0.18	85.56 ± 0.25
DLWCB	64.72 ± 0.22	80.67 ± 0.61	87.02 ± 0.36	70.00 ± 0.34	82.19 ± 0.98	89.96 ± 1.06
MRSCNN	70.67 ± 1.32	81.33 ± 0.04	84.83 ± 0.17	74.00 ± 0.22	82.14 ± 0.27	88.67 ± 0.77
AMARSN	68.73 ± 0.60	78.00 ± 0.12	80.07 ± 0.92	70.03 ± 0.66	81.33 ± 0.11	82.97 ± 0.19
DRSwin-ST	65.55 ± 0.92	74.38 ± 0.75	79.06 ± 0.05	70.04 ± 0.43	78.44 ± 0.44	84.26 ± 0.91
CKG	69.00 ± 0.22	76.12 ± 0.62	81.12 ± 0.27	74.32 ± 0.16	80.39 ± 0.70	86.04 ± 0.74
WD-KANTF	90.50 ± 0.10	97.67 ± 0.18	98.84 ± 0.36	92.36 ± 0.23	98.41 ± 0.02	99.07 ± 0.12

4.4. Ablation experiments

In order to better demonstrate the advantages of AWDL and KANTransformer, 2HP is selected for ablation experiments in Cases A and B. The numbers of training, valid and test samples are 20/20/60. The specific comparison methods of the ablation experiments are as follows:

(1) CNN: CNN consists of four basic convolutional blocks and a fully connected layer. The convolutional block includes convolutional layer, batch normalization layer, activation function and dropout layer. The number of channels is set according to 16, 32, 64, and 128.

(2) CNN + AWDL: CNN + AWDL is an adaptive wavelet denoising layer added after the first convolutional block to verify the denoising effect.

(3) WT-KANTF: The input signals are processed using wavelet transform. Then feature extraction is performed using KANTransformer. This method is used to verify the effect of wavelet transform.

(4) FFT-KANTF: The input signals are processed by using Fast Fourier Transform. Then feature extraction is performed using KANTransformer. This method is used to verify the effect of signal processing.

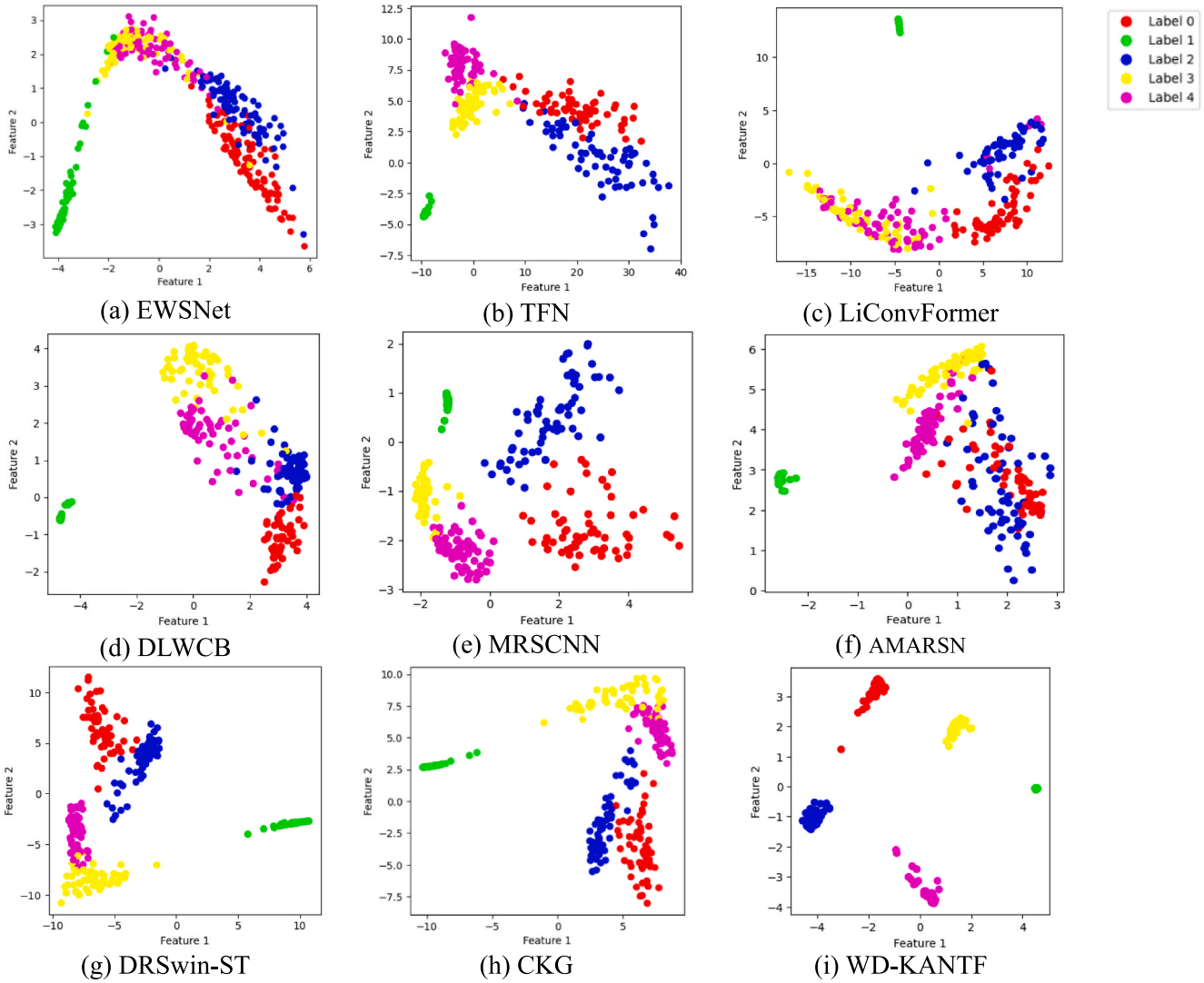


Fig. 10. Visualization results of t-SNE for different methods in Case B.

Table 9
Training time and the number of parameters at 100 epochs.

Condition	Method	Training time	Parameter
Case B	EWSNet	168.22 s	2.89×10^6
3HP	TFN	132.75 s	1.83×10^6
20/20/60	LiConvFormer	84.67 s	3.22×10^5
	DLWCB	176.84 s	1.15×10^6
	MRSCNN	264.13 s	8.97×10^5
	AMARSN	199.72 s	2.43×10^6
	DRSwin-ST	594.75 s	5.24×10^7
	CKG	129.73 s	1.04×10^6
	WD-KANTF	101.25 s	6.41×10^5

(5) CNN + KANTransformer: KANTransformer is used to replace the fully connected layer of CNN in CNN + KANTransformer, which is used to verify the global feature extraction ability and nonlinear feature extraction ability for KANTransformer.

(6) CNN + Transformer: Transformer is used to replace the fully connected layer of CNN in CNN + Transformer, which is used to verify the effect of the designed KANLinear.

(7) WD-KANTF (WDropout): The dropout layer is removed in WD-KANTF.

(8) WD-KANTF (WL2): L2 regularization is removed in WD-KANTF.

Table 10 is the results of the ablation experiments. In Table 10, the diagnostic accuracy of CNN is the lowest. When the adaptive wavelet denoising layer is added into CNN, the diagnostic accuracy of CNN + AWDL increases significantly, which indicates that AWDL can effectively filter out redundant information and improve noise immunity. The diagnostic accuracy of CNN + Transformer is also significantly improved after adding Transformer into CNN, which indicates that it is necessary to perform global feature extraction at the deeper level. Compared with CNN + Transformer, CNN + KANTransformer has better fault diagnosis, which indicates that KANLinear can effectively improve the nonlinear feature extraction ability. Furthermore, the average diagnostic accuracy of WD-KANTF is improved by 6.32 % and 10.31 % compared to WT-KANTF and FFT-KANTF, respectively, which indicates that introducing wavelet knowledge into deep learning model not only improves the ability of capturing critical features, but also enhances the interpretability of the model. Compared with WD-KANTF (WDropout) and WD-KANTF (WL2), the average diagnostic accuracy of WD-KANTF is improved by 3.95 % and 7.46 %, respectively, which indicates that the Dropout layer and L2 regularization can prevent the model from overfitting problem under small sample conditions.

4.5. Model interpretable analysis

4.5.1. Post-hoc explanations of WD-KANTF

Post-hoc explanations can help us further understand the working

Table 10
Fault diagnosis results of ablation experiment (%).

Model	CWRU			RG		
	SNR = −2dB	SNR = 0 dB	SNR = 2 dB	SNR = −2dB	SNR = 0 dB	SNR = 2 dB
CNN	64.67 ± 0.43	73.95 ± 1.02	77.73 ± 0.82	71.09 ± 1.63	75.00 ± 0.02	77.34 ± 0.78
CNN + AWDL	86.15 ± 0.11	92.00 ± 0.21	95.64 ± 0.26	91.38 ± 0.22	93.08 ± 0.54	95.24 ± 0.33
WT-KANTF	84.79 ± 0.33	89.23 ± 0.88	92.17 ± 0.21	86.04 ± 0.97	88.63 ± 1.65	92.79 ± 0.21
FFT-KANTF	80.05 ± 0.17	84.62 ± 0.04	87.77 ± 0.82	83.81 ± 0.06	85.52 ± 0.55	87.92 ± 0.74
CNN + KANTransformer	82.67 ± 1.51	88.28 ± 1.28	92.83 ± 0.57	83.87 ± 1.67	87.33 ± 0.40	90.12 ± 1.06
CNN + Transformer	79.42 ± 0.55	86.05 ± 1.05	88.17 ± 0.96	81.67 ± 0.26	85.16 ± 0.62	88.98 ± 0.02
WD-KANTF (WDropout)	86.46 ± 0.76	92.33 ± 0.07	93.01 ± 0.14	90.75 ± 0.12	92.08 ± 0.33	93.22 ± 0.64
WD-KANTF (WL2)	81.88 ± 0.04	87.76 ± 1.12	88.09 ± 0.98	87.22 ± 0.43	89.36 ± 1.21	92.44 ± 0.75
WD-KANTF	89.22 ± 1.45	95.33 ± 0.13	98.00 ± 0.04	94.33 ± 0.11	96.67 ± 0.03	98.00 ± 0.01

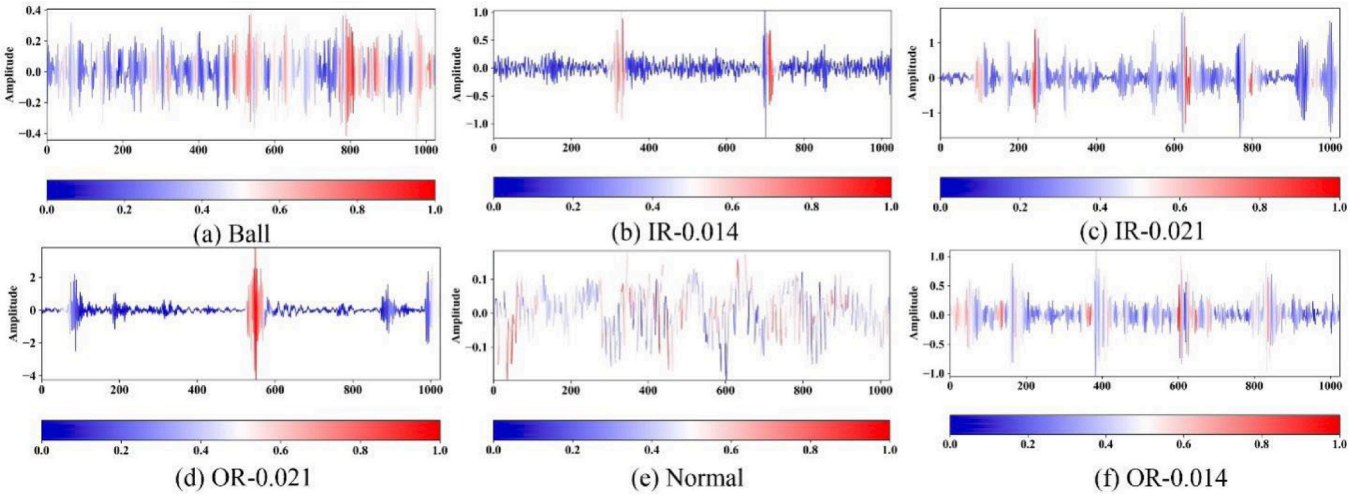


Fig. 11. TGAM visualisation results for Case A.

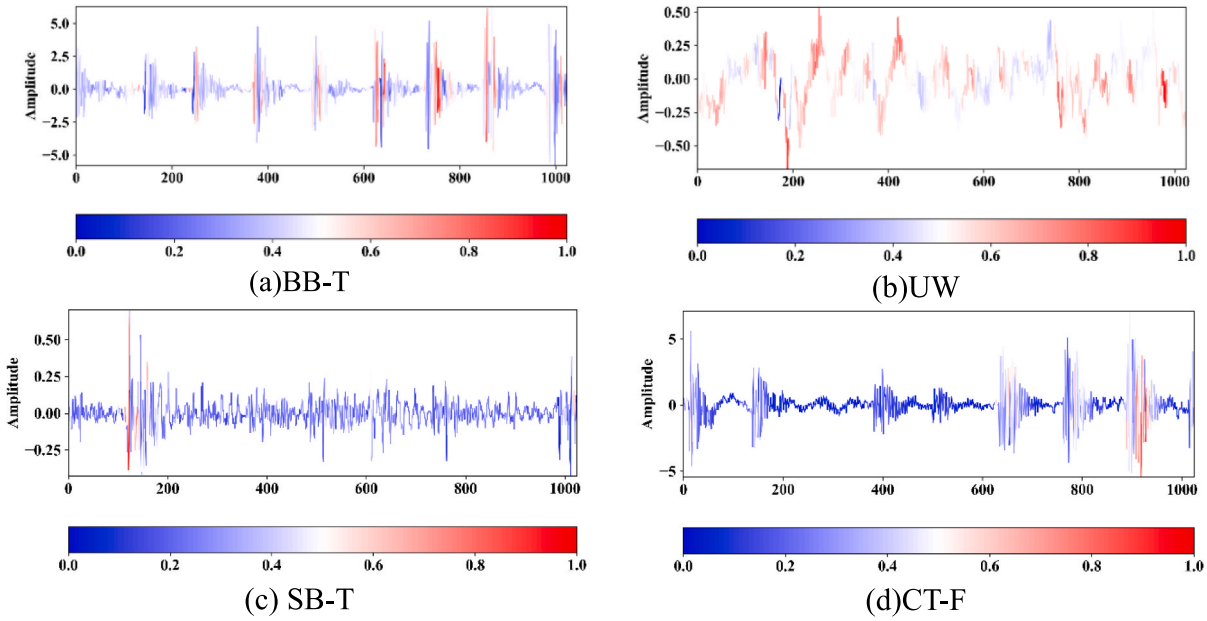
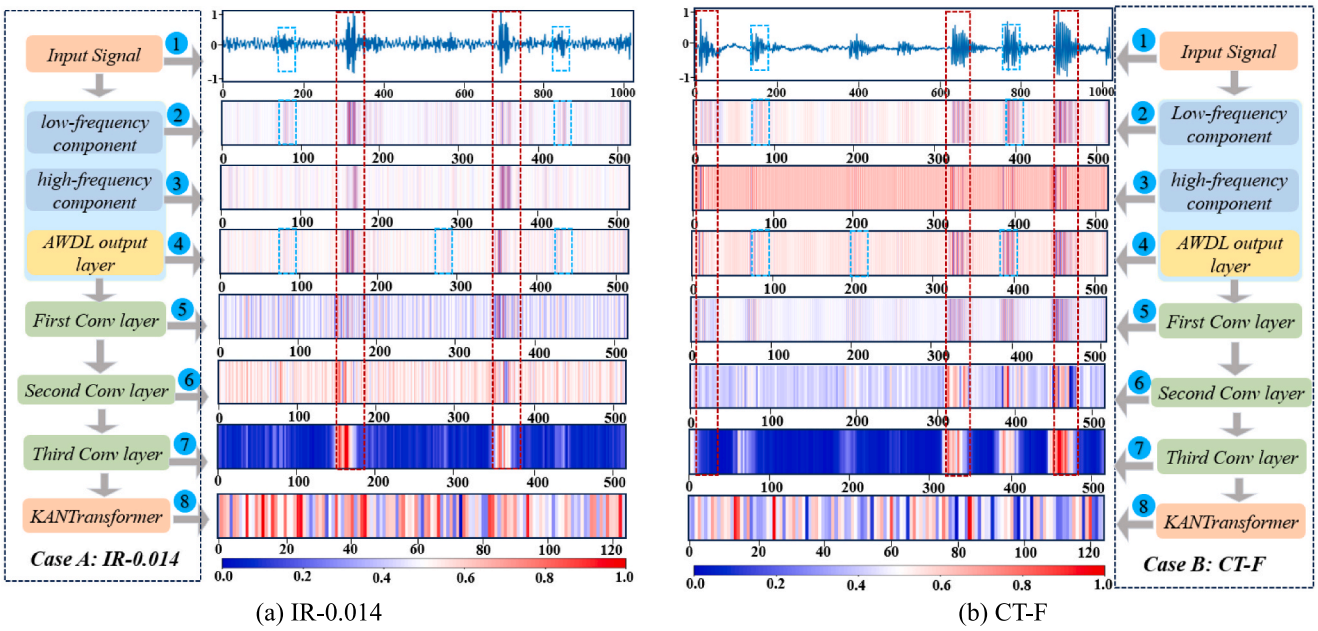


Fig. 12. TGAM visualisation results for Case B.

principle of the model. Therefore, TGAM is used to interpret WD-KANTF. The detailed explanations of TGAM are provided in literature [35]. Fig. 11 and Fig. 12 are the results of the visualizations for Case A and Case B, blue colour indicates that the region receives less attention and red colour indicates that the region receives more attention. In Fig. 11 and Fig. 12, WD-KANTF can focus on shorter pulses, for example, in Fig. 11 (b) and (d), the colour of the pulse region is red, in Fig. 12 (a) and (c), the colour of the shorter pulse region is also red. This shows that TGAM can focus on important pulse features in different datasets. When the amplitude of the vibration signals is stable, WD-KANTF pays more



(a) IR-0.014 (b) CT-F

Fig. 13. Visualisation of feature maps for WD-KANTF from different datasets.

attention to the regions where the rise is obvious, for example, in Fig. 11 (e) and Fig. 12(b), the rising significant amplitude region has the higher attention value and the colour is red, this indicates that WD-KANTF can capture fault features in the input data and gives higher attention values to the regions with fault features.

4.5.2. *WD-KANTF feature learning process*

In order to further explore the feature learning process, the output features of each layer for WD-KANTF are visualized. The results are shown in Fig. 13, which shows the average values of the feature maps at channel index, feature values are represented by the colour mapping, red denotes higher feature values and blue denotes lower feature values. From Fig. 13, we can conclude the followings: (1) WD-KANTF can focus on the impulse region and assign higher feature values, as shown in the red boxes, this shows that WD-KANTF can learn fault features effectively. (2) AWDL prefers to learn the features in the low-frequency components and pays less attention to the features in the high-frequency components, as shown in the blue boxes, this shows that the input signals are converted from the time domain into the wavelet domain, which not only learns more fault features but also enhances the interpretability. (3) The convolutional layer can filter irrelevant information and retain critical fault features, as shown in the third convolutional layer. (4) KANTransformer uses the self-attention mechanism to model the global dependencies in the input signals, the global feature representation is generated by calculating the similarity matrix between different feature vectors, the feature values of different regions show spatial distribution differences in KANTransformer, with the colour alternating from red to blue. This indicates that KANTransformer can accurately capture global fault features and improve nonlinear feature extraction, which helps WD-KANTF to have better robustness under small sample conditions and noise environments.

5. Conclusion

In this paper, a fault diagnosis method based on wavelet denoising and KANTransformer is proposed. In the proposed method, an adaptive wavelet denoising layer is designed, which enhances the interpretability by converting the time-domain signal into the wavelet domain. Meanwhile, a smoothed soft-thresholding and fusion strategy is designed to adaptively focus on the useful features and suppress the redundant information, thus enhancing the robustness of the model. Then, the KANTransformer is designed to capture global fault features while enhancing the nonlinear feature extraction capability. Finally, the WD-KANTF is constructed based on adaptive wavelet denoising layer and KANTransformer. Experimental validation is carried out with bearing and gearbox datasets, and the results show that the average diagnostic accuracy of WD-KANTF is 98.51 %, which is 5.06 % higher than the state-of-the-art method under small sample conditions. Under small sample conditions and noise environments, the average diagnostic accuracy of WD-KANTF is 92.68 %, which is 12.78 % higher than the state-of-the-art method. This indicates that WD-KANTF can extract critical fault features and has good diagnostic performance and robustness.

Future work will investigate cross-domain fault diagnosis to improve the performance of the method for fault diagnosis under different machines and different operating conditions.

CRediT authorship contribution statement

Yazhou Zhang: Writing – original draft, Visualization, Validation, Methodology. **Xiaoqiang Zhao:** Writing – review & editing, Methodology, Funding acquisition. **Zhenrui Peng:** Supervision, Funding acquisition. **Rongrong Xu:** Supervision, Methodology. **Peng Chen:** Supervision, Funding acquisition.

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.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (No.62263021), the College Industrial Support Project of Gansu Province (2023CYZC−24), the Science and Technology Project of Gansu Province (24JRRA172), the Longyuan Young Innovative Talent Team Project of Gansu Province (310100296012), and the Outstanding Postgraduate Innovation Star Project of Gansu Provincial Department of Education (2025CXZX−491)

Data availability

Data will be made available on request.

References

[1] Y. Han, S. Lv, Q. Huang, Y. Zhang, AMCW-DFFNSA: An interpretable deep feature fusion network for noise-robust machinery fault diagnosis, Knowl.-Based Syst. 301 (2024) 112361.

[2] Y. Xiao, H. Shao, M. Feng, T. Han, J. Wan, B. Liu, Towards trustworthy rotating machinery fault diagnosis via attention uncertainty in transformer, J. Manuf. Syst. 70 (2023) 186–201.

[3] G. Chen, G. Tang, Z. Zhu, VKCNN: An interpretable variational kernel convolutional neural network for rolling bearing fault diagnosis, Adv. Eng. Inf. 62 (2024) 102705.

[4] Y. Zhang, X. Zhao, H. Liang, P. Chen, Multiscale dilated convolution and swin-transformer for small sample gearbox fault diagnosis, Appl. Intell. (2024) 1–17.

[5] Y. Xiao, H. Shao, J. Wang, S. Yan, B. Liu, Bayesian variational transformer: A generalizable model for rotating machinery fault diagnosis, Mech. Syst. Sig. Process. 207 (2024) 110936.

[6] J. Xia, Z. Chen, J. Chen, G. He, R. Huang, W. Li, A digital twin-driven approach for partial domain fault diagnosis of rotating machinery, Eng. Appl. Artif. Intel. 131 (2024) 107848.

[7] Z. Xu, K. Zhao, J. Wang, M. Bashir, Physics-informed probabilistic deep network with interpretable mechanism for trustworthy mechanical fault diagnosis, Adv. Eng. Inf. 62 (2024) 102806.

[8] Y. Lu, C. Liang, D. Zhu, Q. Gao, D. Sun, Bearing fault diagnosis using convolutional sparse representation combined with nonlocal similarity, IEEE Sens. J. 23 (6) (2023) 5937–5948.

[9] Z. Li, X. Ding, Z. Song, L. Wang, B. Qin, W. Huang, Digital twin-assisted dual transfer: A novel information-model adaptation method for rolling bearing fault diagnosis, Inf. Fusion 106 (2024) 102271.

[10] W. Li, H. Lan, J. Chen, K. Feng, R. Huang, WavCapsNet: An interpretable intelligent compound fault diagnosis method by backward tracking, IEEE Trans. Instrum. Meas. 72 (2023) 1–11.

[11] H. Liang, J. Cao, X. Zhao, Multi-scale dynamic adaptive residual network for fault diagnosis, Measurement 188 (2022) 110397.

[12] P. Shi, S. Wu, X. Xu, B. Zhang, P. Liang, Z. Qiao, TSN: A novel intelligent fault diagnosis method for bearing with small samples under variable working conditions, Reliab. Eng. Syst. Saf. 240 (2023) 109575.

[13] W. Gong, Y. Wang, M. Zhang, E. Mihankhah, H. Chen, D. Wang, A fast anomaly diagnosis approach based on modified CNN and multisensor data fusion, IEEE Trans. Ind. Electron. 69 (12) (2021) 13636–13646.

[14] Y. Dong, H. Jiang, R. Yao, M. Mu, Q. Yang, Rolling bearing intelligent fault diagnosis towards variable speed and imbalanced samples using multiscale dynamic supervised contrast learning, Reliab. Eng. Syst. Saf. 243 (2024) 109805.

[15] Z. Wu, H. Zhang, J. Guo, Y. Ji, M. Pecht, Imbalanced bearing fault diagnosis under variant working conditions using cost-sensitive deep domain adaptation network, Expert Syst. Appl. 193 (2022) 116459.

[16] Q. Qian, Y. Wang, T. Zhang, Y. Qin, Maximum mean square discrepancy: a new discrepancy representation metric for mechanical fault transfer diagnosis, Knowl.-Based Syst. 276 (2023) 110748.

[17] C. He, H. Shi, X. Liu, J. Li, Interpretable physics-informed domain adaptation paradigm for cross-machine transfer diagnosis, Knowl.-Based Syst. 288 (2024) 111499.

[18] C. Geng, S. Buyun, F. Gaocai, C. Xiangxiang, Z. Guangde, A GAN-based method for diagnosing bodywork spot welding defects in response to small sample condition, Appl. Soft Comput. 157 (2024) 111544.

[19] J. Lin, H. Shao, X. Zhou, B. Cai, B. Liu, Generalized MAML for few-shot cross-domain fault diagnosis of bearing driven by heterogeneous signals, Expert Syst. Appl. 230 (2023) 120696.

[20] Y. Sun, H. Tao, V. Stojanovic, Pseudo-label guided dual classifier domain adversarial network for unsupervised cross-domain fault diagnosis with small samples, Adv. Eng. Inf. 64 (2025) 102986.

[21] H. Liang, J. Cao, X. Zhao, Multibranch and multiscale dynamic convolutional network for small sample fault diagnosis of rotating machinery, IEEE Sens. J. 23 (8) (2023) 8973–8988.

[22] J. Wang, H. Shao, S. Yan, B. Liu, C-ECAFormer: A new lightweight fault diagnosis framework towards heavy noise and small samples, Eng. Appl. Artif. Intel. 126 (2023) 107031.

[23] Y. Sun, H. Tao, and V. Stojanovic, "End-to-end multi-scale residual network with parallel attention mechanism for fault diagnosis under noise and small samples," ISA transactions, 2024.

[24] Q. Chen, et al., TFN: An interpretable neural network with time-frequency transform embedded for intelligent fault diagnosis, Mech. Syst. Sig. Process. 207 (2024) 110952.

[25] Y. Dong, H. Jiang, X. Wang, M. Mu, W. Jiang, An interpretable multiscale lifting wavelet contrast network for planetary gearbox fault diagnosis with small samples, Reliab. Eng. Syst. Saf. 251 (2024) 110404.

[26] S. Li, et al., Digital twin-assisted interpretable transfer learning: a novel wavelet-based framework for intelligent fault diagnostics from simulated domain to real industrial domain, Adv. Eng. Inf. 62 (2024) 102681.

[27] X. Zhang, H. Wang, C. Wang, M. Liu, G. Xu, Time-segment-wise feature fusion transformer for multi-modal fault diagnosis, Eng. Appl. Artif. Intel. 138 (2024) 109358.

[28] Y. Hou, T. Li, J. Wang, J. Ma, Z. Chen, A lightweight transformer based on feature fusion and global-local parallel stacked self-activation unit for bearing fault diagnosis, Measurement (2024) 115068.

[29] F. Lei, Z. Chen, X. Luo, L. Xu, T. Xue, J. Jiang, AHFormer: Hypergraph embedding coding transformer and adaptive aggregation network for intelligent fault diagnosis under noise interference, Adv. Eng. Inf. 61 (2024) 102518.

[30] M.H. Sulaiman, Z. Mustaffa, M.S. Saealal, M.M. Saari, A.Z. Ahmad, Utilizing the Kolmogorov-Arnold Networks for chiller energy consumption prediction in commercial building, J. Build. Eng. 96 (2024) 110475.

[31] M.H. Sulaiman, Z. Mustaffa, A.I. Mohamed, A.S. Samsudin, M.I.M. Rashid, Battery state of charge estimation for electric vehicle using kolmogorov-arnold networks, Energy (2024) 133417.

[32] H. Gao, X. Zhang, X. Gao, F. Li, H. Han, Multi-timescale attention residual shrinkage network with adaptive global-local denoising for rolling-bearing fault diagnosis, Knowl.-Based Syst. 304 (2024) 112478.

[33] D. Sun, Z. Meng, Y. Guan, J. Liu, W. Cao, F. Fan, Intelligent fault diagnosis scheme for rolling bearing based on domain adaptation in one dimensional feature matching, Appl. Soft Comput. 146 (2023) 110669.

[34] Q. Wang, F. Xu, A novel rolling bearing fault diagnosis method based on adaptive denoising convolutional neural network under noise background, Measurement 218 (2023) 113209.

[35] C. He, H. Shi, J. Si, J. Li, Physics-informed interpretable wavelet weight initialization and balanced dynamic adaptive threshold for intelligent fault diagnosis of rolling bearings, J. Manuf. Syst. 70 (2023) 579–592.

[36] X. Zhang, C. He, Y. Lu, B. Chen, L. Zhu, L. Zhang, Fault diagnosis for small samples based on attention mechanism, Measurement 187 (2022) 110242.

[37] S. Yan, H. Shao, J. Wang, X. Zheng, B. Liu, LiConvFormer: A lightweight fault diagnosis framework using separable multiscale convolution and broadcast self-attention, Expert Syst. Appl. 237 (2024) 121338.

[38] X. Zhao, Y. Zhang, An intelligent diagnosis method of rolling bearing based on multi-scale residual shrinkage convolutional neural network, Meas. Sci. Technol. 33 (8) (2022) 085103.

[39] T. Zhou, D. Yao, J. Yang, C. Meng, A. Li, X. Li, DRSwin-ST: an intelligent fault diagnosis framework based on dynamic threshold noise reduction and sparse transformer with shifted windows, Reliab. Eng. Syst. Saf. 250 (2024) 110327.

[40] F. Zhan, L. Hu, W. Huang, Y. Dong, H. He, G. Wu, Category knowledge-guided few-shot bearing fault diagnosis, Eng. Appl. Artif. Intel. 139 (2025) 109489.

Feature and Joint Distribution Migration Alignment Method for Cross-Domain Fault Diagnosis of Rotating Machinery

Yazhou Zhang, Xiaoqiang Zhao, and Rongrong Xu

***Abstract*— Currently, most existing fault diagnosis methods based on domain adaptive (DA) learning reduce the distribution difference between two domains from the metric distance perspective. However, the domain alignment is performed only from the perspective of metric distance without fully mining the transferable features of the samples, which also leads to poor cross-domain fault diagnosis. To address this issue, a domain adversarial fault diagnosis method based on features and joint distribution migration alignment (FJDMA) is proposed. First, a feature aligner is designed to learn more transferable features from both local and global aspects. Second, a new weighted maximum mean square discrepancy (WMMSD) is designed to measure the distribution distance between the samples. The WMMSD can effectively reduce the distribution distance between the same classes within the domain. In addition, to increase the distribution distance between different classes between domains, we introduce correlation alignment (CORAL). Finally, a dynamic factor is designed to quantitatively combine WMMSD and CORAL, thus constructing the joint distribution migration (JDM). The JDM further enhances domain confusion during model training. Two bearing datasets and one gear dataset are used for experimental validation. The results show that the average diagnostic accuracy of the proposed method between the two bearing datasets is 3.34% higher than that of the state-of-the-art method. The average diagnostic accuracy of the proposed method between the bearing and gear datasets is improved by 2.38% over the state-of-the-art method.**

***Index Terms*— Attention mechanism, domain adaptation, fault diagnosis, global alignment, joint distribution migration (JDM).**

I. Introduction

ROTATING machinery is widely used in the fields of energy, transport, and manufacturing. In recent years, with the continuous progress of technology, the research on rotating machinery has achieved rich results. For example, Zhang et al. [1], [2] investigated the application of rotating machinery in wind turbines. Lu et al. [3] explored the challenges of rotating machinery in high-speed trains. However, there are complex interactions among the various components of rotating equipment. A partial failure may cause the whole rotating equipment to stop working and even cause casualty accidents [4], [5]. Bearings and gears are the key components of rotating equipment. Since they often work in harsh environments, this leads to their vulnerability to faults. Therefore, in order to detect and eliminate faults as early as possible, it is essential to carry out fault diagnosis for gears and bearings [6], [7], [8].

Fault diagnosis methods for rotating equipment have gradually developed from relying on expert experience and knowledge to using data-driven fault identification. Currently, with the development of sensors and computer technologies, data-driven fault diagnosis methods have further received increasing attention. In particular, deep learning is highly popular by Zhao and Shen [9] and Wu et al. [10]. Deep learning can automatically extract the features from input signals, thus reducing the dependence on professional knowledge and the influence of human factors on the feature extraction process. Therefore, deep learning-based fault diagnosis methods are widely used in the field of fault diagnosis [11], [12], [13]. For example, Yan et al. [14] combined broadcast self-attention with separable convolution to construct a lightweight model of LiConvFormer for fault diagnosis. Zhao and Zhang [15] improved the residual contraction network by multiscale technique and proposed a multi-scale residual shrinkage convolutional neural network (MRSCNN) bearing fault diagnosis method to verify its superiority. The above methods are based on two premises: 1) sufficient labeled data and 2) the divided training samples and test samples have the same distribution. However, in real industry, the working conditions of rotating equipment are complex and variable. Replacement of components, speed, and load changes can lead to the existence of distribution differences in the data, which would cause serious degradation of the performance of deep learning-based fault diagnosis methods.

The development of domain adaptive (DA) provides new ideas to solve the above problems. It performs fault identification by mining the domain-invariant features between the source and target domains. Specifically, DA extracts the fault information through a neural network and inputs the fault information into the adaptive layer to reduce the distribution distance between the source and target domains [16], [17], the process of which is illustrated in Fig. 1(a).

Received 15 August 2024; revised 27 October 2024; accepted 19 November 2024. Date of publication 24 March 2025; date of current version 4 April 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62263021 and in part by the Industrial Support Project of Education Department of Gansu Province under Grant 2021CYZC-02. The Associate Editor coordinating the review process was Dr. Ferdinanda Ponci. *(Corresponding author: Xiaoqiang Zhao.)*

The authors are with the School of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou, Gansu 730050, China (e-mail: 1911599612@qq.com; xqzhao@lut.edu.cn; 2313875156@qq.com).

Digital Object Identifier 10.1109/TIM.2025.3548217

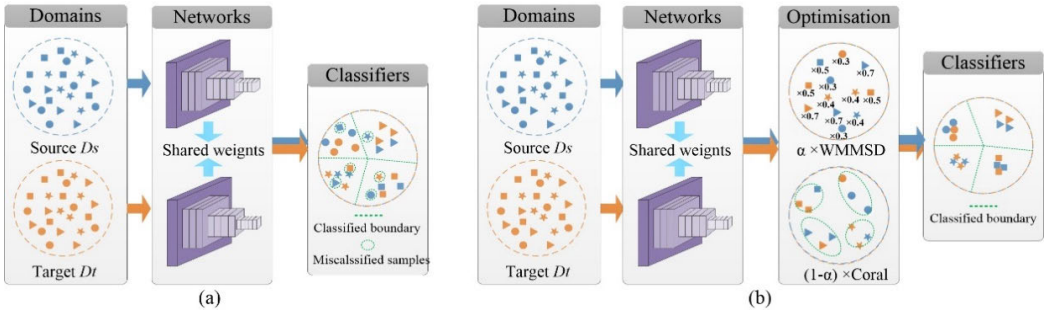


Fig. 1. Results of minimizing MMD and JDM are shown in (a) and (b), respectively.

For example, Xiao et al. [18] proposed a domain adaptive method of multiscale convolutional subdomain adaptation network, which evaluated the difference between the relative subdomains of two domains by using the local maximum mean difference. Han et al. [19] proposed a dual classifier correlation transfer diagnostic method, which achieved global and local feature alignment by using maximum mean difference and local maximum mean difference. In addition, to better reduce domain distributional differences, some scholars have constructed joint distributional discrepancies to further reduce the distributional distance between the two domains. Specifically, the class conditional probabilities are used to approximate the class posterior probabilities, and the marginal and conditional distributions of the two domains are aligned simultaneously. For example, Liu et al. [20] proposed a duplex adversarial deep discriminative network (DADDN), which constructed the weighted joint distributions using conditional maximum mean discrepancy (MMD) measurement functions to align the differences between the source and target domains. Sun et al. [21] proposed a joint discriminative adversarial domain adaptive method, which designed a discriminative discrepancy module to achieve class alignment and combines class alignment with domain alignment, which ultimately achieves fault diagnosis under cross-domain.

The above studies only reduce the domain differences by measuring the distance between two domain distributions, without considering whether the extracted fault information by the neural network is suitable for feature transfer. The attention mechanism can help models acquire useful features and suppress redundant features and is widely used in deep learning models [22], [23], [24]. Therefore, some scholars have used the attention mechanism to improve the performance of domain adaptive fault diagnosis. For example, Chen et al. [25] proposed a joint-attention adversarial domain adaptive approach, which combined MMD with an attention mechanism to enhance transferable features by attentively weighting the maximum mean discrepancy processed features. Shao and Kim [26] proposed an adaptive multiscale attention convolutional network for cross-domain fault detection, which used multiple attention mechanisms to obtain more domain-invariant features in the feature extraction module. In addition, to refine the domain-invariant features and pass them to higher space, Yao et al. [27] proposed a hierarchical adversarial domain adaptive method, which used a multistate attention mechanism and a coding network to reduce the distributional difference between two domains, and achieved gear fault diagnosis through a variable feature mining strategy.

In summary, researchers have investigated transfer learning methods from several aspects and achieved high diagnostic accuracy in the field of fault diagnosis. However, the operating environment of rotating equipment is complex and variable, and it is often subject to strong noise and harmonic interference, resulting in the swamping of valuable features. Therefore, considering only domain alignment or feature alignment leads to insufficient domain adaptivity, which reduces the diagnosis accuracy. To address this issue, this article proposes a domain adaptive method based on feature and joint distribution migration alignment (FJDMA). Existing domain adaptive methods mostly focus on the aspect of metric distance to reduce the distribution difference between two domains. However, the domain alignment is only from metric distance without fully exploiting the transferable features of the samples, which also leads to poor cross-domain diagnosis. The FJDMA performs domain adaptation from both feature and metric distance aspects. At the feature aspect, the attention mechanism is used to focus on more domain-invariant features and suppress redundant features. At the aspect of metric distance, a joint distribution migration (JDM) loss is designed to help the model further reduce the two-domain distribution difference during the training process. Specifically, first the method utilizes a feature extractor to extract fault features between the two domains. Then, a feature alignment module is designed, which focuses on more transferable features and suppresses nontransferable features at both local and global aspects. In the feature alignment module, we use the local alignment module to refine the domain-invariant features at the channel and spatial level for the low-dimensional feature space. In the high-dimensional feature space, we design the fully connected self-attention (FCSA) mechanism that helps the model acquire useful fault features and suppresses the effect of redundant features. In addition, a JDM loss is designed to further reduce the domain differences during model domain adversarial training. Fig. 1(b) shows an illustrative example to show the effectiveness of JDM. The main contributions of this article are as follows.

1) A feature aligner is designed, which can weigh the extracted features of the feature extractor both locally and globally to enhance the transferable features and provide quality domain-invariant features for the classifier.

2) A JDM loss is constructed to reduce domain differences during model training. Specifically, to measure the intradomain class distribution distance from the mean square value perspective, a weighted maximum mean square discrepancy (WMMSD) loss is designed. In addition, to measure the interdomain class distribution distance from the covariance perspective, a correlation alignment (CORAL) loss is introduced. Finally, a dynamic factor is designed to quantitatively assign the weights to WMMSD and CORAL.
3) The experimental validation is performed using three datasets. Specifically, the cross-domain experiments are conducted by adding noise to the target domain in a single dataset; the experiments are conducted in multiple datasets of different subjects. The results exhibit that the proposed method has excellent diagnostic performance.

The rest of this article is organized as follows. Section II introduces the theoretical knowledge of domain adversarial networks. Section III describes the proposed method in detail. Section IV performs the experimental validation and analysis, and finally, Section V concludes this article.

II. Guidelines for Manuscript Preparation

A. Problem Definition

This article investigates the problem of domain adaptation in fault diagnosis. First, we define some notations and basic concepts. Assume that a labeled single source domain is $D_s = \{X_s, Y_s\} = \{(x_i, y_i)\}_{i=1}^{N_s}$, where X_s and Y_s denote the data samples and health state labels, respectively. N_s denotes the number of source domain samples. An unlabelled target domain denotes $D_t = \{X_t\} = \{(x_i)\}_{i=1}^{N_t}$, and N_t denotes the number of target domain samples.

The source and target domains share the same label space in domain adaptation, i.e., $Y_s = Y_t$. However, the marginal probability distributions and conditional probability distributions of the two domain samples are different in the working environment. The marginal probability distribution is $P_s(X_s) \neq P_t(X_t)$ and the conditional probability distribution is $Q_s(Y_s|X_s) \neq Q_t(Y_t|X_t)$. This results that the generalization performance of transfer learning is degraded between the source and target domains. Therefore, the goal of this article is that an intelligent diagnostic model is developed to effectively reduce the differences in the marginal and conditional probability distributions between the source and target domains and to improve the generalization performance of the model.

B. Domain Adversarial Neural Network (DANN)

DANN and its variants are widely used in transfer learning tasks, which reduce the data distribution discrepancy between the source and target domains by adversarial training [28], [29], [30]. It consists of a feature extractor G_f, domain discriminator D_d, and classifier C_y. The domain discriminator is unique to domain adversarial networks. During its training, a gradient inversion layer is added to the classifier. The optimization objective of DANN is to minimize the classification loss and task loss. The classification loss is achieved by using the domain discriminator, which aims to reduce the differences in the marginal probability distributions between the two domains. The task loss is used to minimize the prediction error by the classifier to achieve the accurate prediction of the samples. The optimization objective of network can be defined as follows:

$$L(\vartheta_f, \vartheta_y, \vartheta_d) = \frac{1}{N_s} \sum_{x_i \in X_s} L_y(C_y(G_f(x_i)), y_i) - \frac{\lambda}{N_s + N_t} \sum_{x_i \in X_s \cup X_t} L_d(D_d(C_y(x_i)), d_i) \quad (1)$$

where ϑ_f, ϑ_y, and ϑ_d denote the parameters of G_f, C_y, and D_d, respectively; y_i and d_t denote the data sample labels and domain labels, respectively; $L(\cdot), L_y(\cdot)$, and $L_d(\cdot)$ denote the total loss, task loss, and domain classification loss, respectively; and λ denotes the hyperparameter between losses.

During DANN training, task loss and classification loss are minimized to improve the predictive power. The adversarial loss between the feature extractor and domain discriminator is maximized to reduce the marginal probability distribution discrepancy. Therefore, the parameters are updated as follows:

$$(\hat{\vartheta}_f, \hat{\vartheta}_y) = \arg\min_{\vartheta_f, \vartheta_y} L(\vartheta_f, \vartheta_y, \hat{\vartheta}_d) \quad (2)$$

$$(\hat{\vartheta}_d) = \arg\max_{\vartheta_d} L(\hat{\vartheta}_f, \hat{\vartheta}_y, \vartheta_d) \quad (3)$$

where $\widehat{\vartheta}_f, \widehat{\vartheta}_y$, and $\widehat{\vartheta}_d$ denote the optimal parameters. The domain adversarial network extracts domain-invariant features by reducing the discrepancy in data distribution between different domains, which further improves the performance of the model on the diagnostic task.

III. Math

The proposed FJDMA aims to address the problem of poor transfer performance from a single source domain to a target domain due to inconsistencies in class and domain features. Its general framework is shown in Fig. 2, and the specific network parameters are shown in Table I. The FJDMA consists of a feature extractor, a feature aligner, a classifier, and a domain discriminator. The specific details of each component are described as follows. In Table I, the local alignment module focuses on more domain-invariant features from the low-dimensional feature space. Specifically, the local alignment module refines the fault features at channel and spatial indexes to extract more domain-invariant features. In addition, we select the domain discriminator with a dropout rate of 0.5 by cross validation in the experimental section, which balances the diagnostic tasks between the source and target domains during the training process, thus effectively avoiding model overfitting.

A. Network Structure

1) Feature Extractor: The 1-D convolutional neural networks can automatically extract deeper and better feature representations from input samples. Therefore, we take advantage of its superior feature extraction capability to

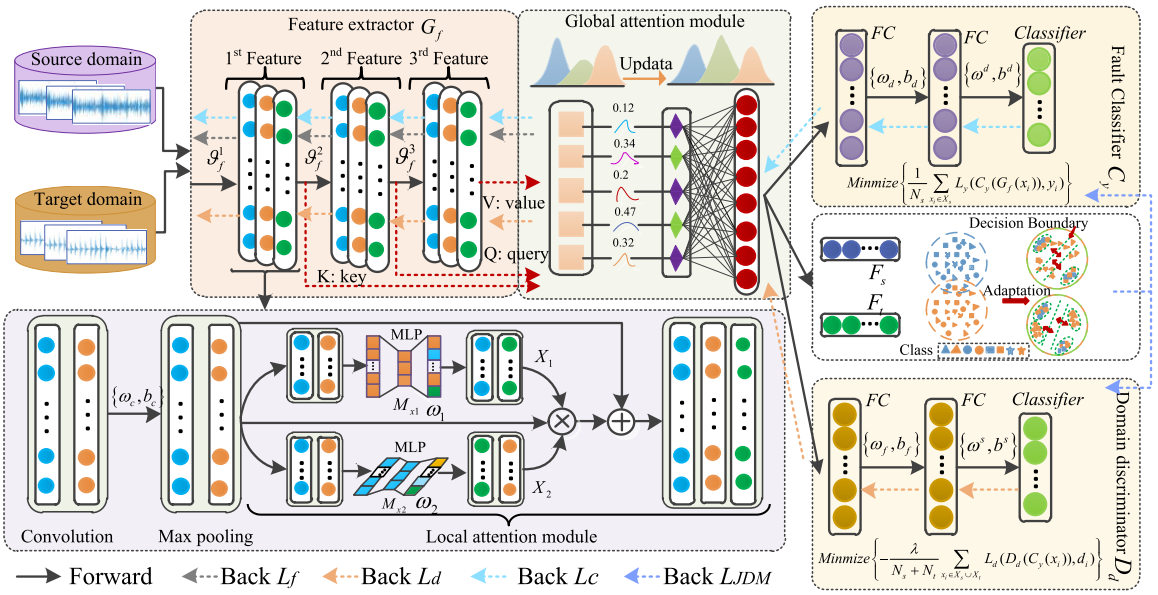


Fig. 2. Overall diagnostic flowchart of the FJDMA.

TABLE I
PARAMETERS OF THE FJDMA

Names	Layers	Operators	Stride / Size	Output
Feature extractor	1st Feature	Convolution	2 / 3	(1024,16)
		BN / ReLu	-	
		Max-pooling	2 / 2	(512,16)
		Local alignment	-	
	2nd Feature	Convolution	2 / 3	(256,64)
		BN / ReLu	-	
		Max-pooling	2 / 2	(128,64)
		Local alignment	-	
	3rd Feature	Convolution	2 / 3	(64,128)
		BN / ReLu	-	
		Max-pooling	2 / 2	(32,128)
		Local alignment	-	
Feature Aligner	global alignment	FC Self-attention mechanism	-	128
Domain discriminator	FC1	Dense (ReLu)	-	128
		Dropout (0.5)	-	
	FC2	Dense (ReLu)		64
		Dropout (0.5)	-	
		Sigmoid	-	1
Classifier	FC3	Dense (ReLu)	-	128
	FC4	Dense (ReLu)	-	64
		Softmax	-	Class

construct the feature extractor. Specifically, the feature extractor consists of three feature extraction modules. Each feature extraction module consists of a convolutional layer, a pooling layer, and a local alignment module.

First, we assume that the sample space formed by the source and target domain samples is $D = \{D_s, D_t\} \in \mathbb{R}^{\{N_s, N_t\}}$, and the convolution operation for the input samples can be expressed as follows:

$$Z_i^l = D_i^{l-1} \times \omega_i^{l-1} + b_i^{l-1} \tag{4}$$

where D_i^{l-1} is the input sample, l denotes the number of convolutional layers, ω_i^{l-1} denotes the weight of the convolutional layers, and b_i^{l-1} denotes the convolutional layer bias. To improve the model training speed and mitigate the vanishing gradient, we use a linear correction unit after the convolutional layers. The process can be expressed as

$$\bar{Z}_i^l = f_{\text{relu}}\left(0, Z_i^l\right) \tag{5}$$

where $f_{\text{relu}}(\cdot)$ denotes the ReLu activation function.

Second, we add the output to the maximum pooling (Max-pooling) layer. The Maxpooling reduces the dimensionality of the feature map and optimizes the parameters. The process can be described as follows:

$$Z\left(\text{Pool}_i^l\right) = \max\left(\bar{Z}_i^l, s\right) \tag{6}$$

where s denotes the parameter that regulates the size of the output feature.

Finally, we input the output feature maps to the local alignment module. The local alignment module consists of channel attention and spatial attention in residual parallelism. Spatial attention is responsible for extracting the spatial features of the input samples, while channel attention can retain the extracted detailed features. Then, the processed features are fused. The process of spatial attention calculation can be described as follows:

$$\begin{cases} A^{\text{sp}}(z) = \omega(z) \times f_{\text{st}}[\alpha(\omega_1(z) \times \omega_2(z))] \\ Z^{\text{sp}} = A^{\text{sp}}(z) \otimes_{\text{sp}} z \end{cases} \tag{7}$$

where z denotes the input sample, ω denotes the convolution operation, ω_1 and ω_2 denote the convolution layers, f_{st} denotes the softmax function, and $\otimes_{\text{sp}}$ denotes the spatial multiplication operation. The process of channel attention calculation can be described as follows:

$$\begin{cases} A^{\text{ch}}(z) = z \times f_{\text{st}}\left[\omega(\omega_{\text{map}}(z))\right] \\ Z^{\text{ch}} = A^{\text{ch}}(z) \otimes_{\text{ch}} z \end{cases} \tag{8}$$

where ω_{map} denotes the average pooling operation and $\otimes_{\text{ch}}$ denotes the channel multiplication operation. The output of

the local attention module is described as

$$Z = \left(Z^{\mathrm{sp}} \otimes Z^{\mathrm{ch}}\right) \oplus Z\left(\mathrm{Pool}_i^l\right) \tag{9}$$

where $\otimes$ denotes the dot product operation and $\oplus$ denotes the sum operation.

2) Feature Aligner: Feature extractor can extract the features from original data. However, not all fault features are suitable for transfer learning. If the features that are not applicable to the transfer task are forced to be adapted, it may lead to the occurrence of negative transfer. Therefore, a feature aligner is constructed to adaptively enhance transferable features and suppress nontransferable features in this article. Specifically, the feature aligner consists of a local alignment module and a global alignment module. In the low-dimensional feature space, we use the local alignment module to refine domain-invariant features at the channel and spatial levels. The local alignment module is described in Section III-A. With increasing the number of network layers, the feature extractor can capture richer feature information but may also introduce more redundant information. This may adversely affect the transfer learning performance of the model. Therefore, in the high-dimensional feature space (FC layer), we design the FCSA mechanism to help the network capture useful fault features while suppressing the effect of redundant features. The FCSA is shown in Fig. 3. In Fig. 3, the FCSA processing can be divided into three steps. First, the input features are linearly transformed by the dense layer to obtain the query matrix Z_{q}, key matrix Z_{k}, and value matrix Z_{v}, which are described as follows:

$$\begin{cases} Z_{\mathrm{q}} = X^{\mathrm{input}} \times W_i^Q \\ Z_{\mathrm{k}} = X^{\mathrm{input}} \times W_i^K \\ Z_{\mathrm{v}} = X^{\mathrm{input}} \times W_i^V \end{cases} \tag{10}$$

where W_i^Q, W_i^K, and W_i^V denote the parameter matrices, which are continuously learned by training the network. Second, the key matrix is transposed to obtain $Z_{\mathrm{k}}^{\mathrm{T}}$. $Z_{\mathrm{k}}^{\mathrm{T}}$ and the query matrix are multiplied to obtain the correlation score between the two vectors. This is normalized by the softmax function to obtain the attention weights for the key and query matrices. The process is described as follows:

$$\mathrm{att}_i^{\mathrm{q,k}} = \mathrm{softmax}\left(Z_{\mathrm{q}} \times Z_{\mathrm{k}}^{\mathrm{T}} / \sqrt{d}\right) \tag{11}$$

where d is the dimension of the parameter matrix. Finally, the value matrix and the attention weights perform a dot product operation to obtain the output of FCSA. The process is described as follows:

$$X^{\mathrm{output}} = Z_{\mathrm{v}} \otimes \mathrm{att}_i^{\mathrm{q,k}}. \tag{12}$$

In high-dimensional feature space, domain-invariant features are further enhanced and the distribution between source and target domains is reduced by FCSA.

3) Classifiers: In DA, the classifier aims to map the fault feature information to the corresponding class labels. It consists of three FC layers and a dropout. The softmax function is used to estimate the class probability for the fault diagnosis in the last FC layer.

4) Domain Discriminator: The domain discriminator is designed to identify the source of the input data and thus determine whether the data come from a known source domain or an unknown target domain. It consists of three FC layers. In contrast to the classifier, the last layer of the domain discriminator uses the sigmoid function to determine the domain type of the samples.

B. Optimized Processing

1) Classifier Loss: The classifier not only classifies the samples in the target domain but also helps the feature extractor to learn domain-invariant features during backward transmission of errors. The loss function of the classifier is the cross-entropy loss function, which is calculated as follows:

$$L_{\mathrm{c}} = -\frac{1}{n_{\mathrm{s}}} \sum_{i=1}^{n_{\mathrm{s}}} y_i \log\left(C_y\left(f_i^s, \theta_{\mathrm{c}}\right)\right) \tag{13}$$

where n_{s} denotes the number of samples from the source domain and y_i denotes the labeling of the source domain samples.

2) Domain Discriminator Loss: In adversarial training, the discriminator and the classifier fight against each other. The classifier tries to deceive the domain discriminator, while the domain discriminator endeavors to distinguish the source of the generated data. Thus, the domain discriminator is a binary classifier, which is computed as follows:

$$L_{\mathrm{d}} = l_{\mathrm{d}} \sum_{i=1}^{n_{\mathrm{s}}} \log\left[D_{\mathrm{d}}\left(f_{\mathrm{s}}^i, \theta\right)\right] + (1 - l_{\mathrm{d}}) \sum_{i=1}^{n_{\mathrm{t}}} \log\left[D_{\mathrm{d}}\left(f_{\mathrm{t}}^i, \theta\right)\right] \tag{14}$$

where l_{d} denotes the domain label 0 or 1, $D_{\mathrm{d}}(f_{\mathrm{s}}^i, \theta)$ denotes the predicted label of the source domain, and $D_{\mathrm{d}}(f_{\mathrm{t}}^i, \theta)$ denotes the predicted label of the target domain.

3) JDM Loss: In DA, the feature aligner can only filter domain-invariant features from the extracted features. However, when facing large differences in working conditions or different datasets, transfer learning is insufficient by feature alignment. Therefore, we consider to further reduce the impact of domain differences from the perspective of model training. Domain differences are usually represented by the increasing distance of the same classes within the domain and the confusion of different classes between domains. Specifically, we define the differences between the source and target domain classes as follows:

$$D_{\mathrm{dis}}(X^{\{s,t\}}) = \left| \begin{array}{l} \sum_{\mathrm{c}=1}^{C} p_{\mathrm{s}}(y^s = c)\, p_{\mathrm{s}}\left(x_{\mathrm{s}}^i \middle| y^s = c\right) \\ -\sum_{\mathrm{c}=1}^{C} p_{\mathrm{t}}(y^t = c)\, p_{\mathrm{t}}\left(x_{\mathrm{t}}^i \middle| y^t = c\right) \end{array} \right| \tag{15}$$

where C denotes the number of classes, $P_{\mathrm{s}}(y^s = c)$ denotes the probability that the source domain sample x_{s}^i denotes categorized as c (class weight), and $P_{\mathrm{t}}(y^t = c)$ denotes the probability that the target domain sample x_{t}^i denotes categorized as c. It can be seen from (15) that we can reduce the distance of the same classes in both domains by

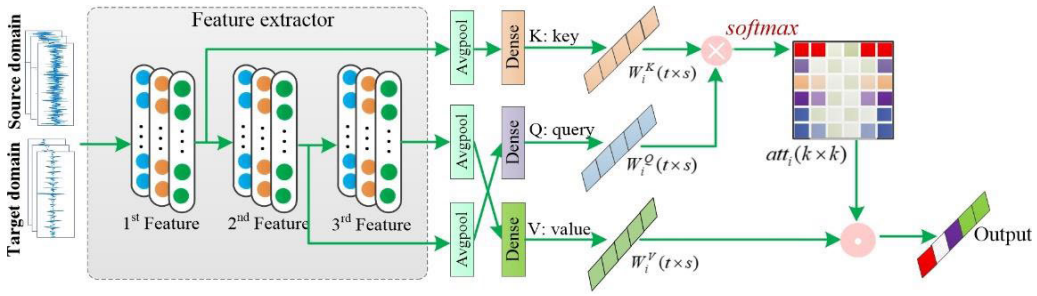


Fig. 3. Schematic of FCSA structure.

minimizing the value of $D_{\text{dis}}(\cdot)$. However, since the label $(p_{\text{t}}(y^t = c))$ of the target domain is unknown in DA, it becomes impractical to use the $D_{\text{dis}}(\cdot)$ measure of difference.

Qian et al. [31] proposed a maximum mean square discrepancy (MMSD) to measure the difference between two domains. Therefore, we use MMSD to calculate the difference between two domains. However, this metric assigns different weights to domain classes, resulting in larger distances between the same classes. Therefore, we design a WMMSD. First, MMSD is defined as follows:

$$\begin{aligned}\text{MMSD}^2 = &\frac{1}{n_{\text{s}}^2}\sum_{i=1}^{n_{\text{s}}}\sum_{j=1}^{n_{\text{s}}}\left[k\left(x_{\text{s}}^i\right), k\left(x_{\text{s}}^j\right)\right]^2 \\ &+ \frac{1}{n_{\text{t}}^2}\sum_{i=1}^{n_{\text{t}}}\sum_{j=1}^{n_{\text{t}}}\left[k\left(x_{\text{t}}^i\right), k\left(x_{\text{t}}^j\right)\right]^2 \\ &- \frac{2}{n_{\text{s}}n_{\text{t}}}\sum_{i=1}^{n_{\text{s}}}\sum_{j=1}^{n_{\text{t}}}\left[k\left(x_{\text{s}}^i\right), k\left(x_{\text{t}}^j\right)\right]^2 \end{aligned} \tag{16}$$

where n_{s} and n_{t} denote the number of samples in the source and target domains, respectively; $k(\cdot)$ denotes the regeneration kernel function; x_{s}^i denotes the ith sample in the source domain; and x_{t}^j denotes the jth sample in the target domain. Second, to achieve the source and target domains having the same class weight, we define the class weight factor as follows:

$$\omega = \frac{\omega_{\text{s}}^c - \omega_{\text{t}}^c}{\omega_{\text{s}}^c} = \frac{p_{\text{s}}(y^s = c) - p_{\text{t}}(y^t = c)}{p_{\text{s}}(y^s = c)} \tag{17}$$

where ω_{s}^c is the probability that the source domain samples are classified as class c and ω_{t}^c is the probability that the target domain samples are classified as class c. Since ω_{t}^c obtains from the pseudo-labeling of the target domain, the WMMSD can be described as follows:

$$\begin{aligned}\text{WMMSD}^2 = &\frac{1}{\left(\sum_{i,j}^{n_{\text{s}}}\omega_{i,j}\right)^2}\sum_{i=1}^{n_{\text{s}}}\sum_{j=1}^{n_{\text{s}}}\omega_{i,j}\cdot\left[k\left(x_{\text{s}}^i\right), k\left(x_{\text{s}}^j\right)\right]^2 \\ &+ \frac{1}{n_{\text{t}}^2}\sum_{i=1}^{n_{\text{t}}}\sum_{j=1}^{n_{\text{t}}}\left[k\left(x_{\text{t}}^i\right), k\left(x_{\text{t}}^j\right)\right]^2 \\ &- \frac{2}{n_{\text{t}}\cdot\sum_i^{n_{\text{s}}}\omega_i}\sum_{i=1}^{n_{\text{s}}}\sum_{j=1}^{n_{\text{t}}}\omega_i\cdot\left[k\left(x_{\text{s}}^i\right), k\left(x_{\text{t}}^j\right)\right]^2 \end{aligned} \tag{18}$$

where $\sum_{i,j}^{n_{\text{s}}}\omega_{i,j}$ is the average of the class weights of the source domain samples. It can be seen from (18) that we can be seen that the class weights are computed by averaging over all the pseudo-labels, thus ensuring the robustness of the metric. Finally, the WMMSD loss can be obtained from (18) as follows:

$$L_{\text{wmmsd}} = \left\|E_p[k(X_{\text{s}})\otimes k(X_{\text{s}})] - E_{\text{q}}[k(X_{\text{t}})\otimes k(X_{\text{t}})]\right\|_{H\otimes H}^2 \tag{19}$$

where H is the regenerated kernel Hilbert space; $k(\cdot)$ is the regenerated kernel function; X_{s} and X_{t} are the set of samples in the source and target domains, respectively; and $\otimes$ is the mean square computation procedure.

WMMSD loss enhances the competitiveness between classes in the intradomain by assigning the same weights to source and target domain classes. However, it is still necessary that the distances between different classes in the interdomain need to deserve attention during domain adaptation learning. The CORAL loss can solve the problem well. Therefore, we use CORAL to reduce the differences in feature distributions between interdomain classes. CORAL loss is calculated as follows:

$$\begin{cases}\text{Mean}_{\text{s}} = \dfrac{1}{n_{\text{s}}}\displaystyle\sum_{i=1}^{n_{\text{s}}}x_{\text{s}}^i \\ \text{Mean}_{\text{t}} = \dfrac{1}{n_{\text{t}}}\displaystyle\sum_{i=1}^{n_{\text{t}}}x_{\text{t}}^i\end{cases} \tag{20}$$

$$\begin{cases}\text{Cov}_{\text{s}} = \dfrac{1}{n_{\text{s}}-1}\displaystyle\sum_{i=1}^{n_{\text{s}}}\left(x_{\text{s}}^i - \frac{1}{n_{\text{s}}}\sum_{i=1}^{n_{\text{s}}}x_{\text{s}}^i\right)^{\text{T}}\left(x_{\text{s}}^i - \frac{1}{n_{\text{s}}}\sum_{i=1}^{n_{\text{s}}}x_{\text{s}}^i\right)\right) \\ \text{Cov}_{\text{t}} = -\dfrac{1}{n_{\text{t}}-1}\displaystyle\sum_{i=1}^{n_{\text{t}}}\left(x_{\text{t}}^i - \frac{1}{n_{\text{t}}}\sum_{i=1}^{n_{\text{t}}}x_{\text{t}}^i\right)^{\text{T}}\left(x_{\text{t}}^i - \frac{1}{n_{\text{t}}}\sum_{i=1}^{n_{\text{t}}}x_{\text{t}}^i\right)\right)\end{cases} \tag{21}$$

$$L_{\text{coral}} = \|\text{Mean}_{\text{s}} - \text{Mean}_{\text{t}}\|^2 - \|\text{Cov}_{\text{s}} - \text{Cov}_{\text{t}}\|_{\text{F}}^2 \tag{22}$$

where Mean_{s} is the source domain mean, Mean_{t} is the target domain mean, Cov_{s} is the source domain covariance matrix, Cov_{t} is the target domain covariance matrix, $\|\cdot\|^2$ is the square of the Euclidean distance, and $\|\cdot\|_{\text{F}}^2$ is the square of Frobenius.

The WMMSD loss and CORAL loss can address domain differences in model training. However, how to quantitatively measure the role of the two losses is crucial to improve the performance of model transfer learning in model training. Therefore, we design a dynamic adaptive factor to quantitatively assign weights to the WMMSD loss and CORAL loss. This dynamic factor can be viewed, as shown

Algorithm 1 Training and Test Procedures for FJDMA

Inputs: source domain labeled dataset $D_s = \{(x_i, y_i)\}_{i=1}^{N_s}$, target domain unlabelled dataset $D_t = \{(x_i)\}_{i=1}^{N_t}$

Set hyperparameters such as convolution layer, activation function, pooling layer, learning rate, batch size, epoch
To initialize the weights and biases of the model
Divide the training and test sets of the source and target domains
For each training epoch **do**
  If Phase = train **then**
    For each batch **do**
      Calculation the output of C_y
      Solve L_c based on Eq. (13)
      Calculation the output of D_d and JDM
      Solve L_d based on Eq. (14)
      Solve L_{wmmsd} and L_{coral} based on Eq. (19) and (22)
      Solve the dynamic factor α f based on Eq. (23) and (24)
      Calculate the output of L_{JDM} from Eq. (25)
      Obtain the optimization objective function L_{total} from Eq. (26)
      Update model parameters
    End for
  End For
  Save the trained model

Output: Test sets is loaded into the trained model to obtain diagnostic results

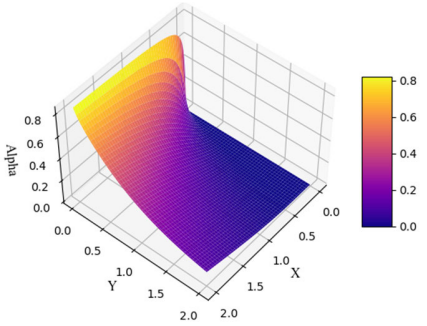


Fig. 4. Visualization of dynamic factors.

in Fig. 4. In Fig. 4, the X-axis is the WMMSD loss and the Y-axis is the CORAL loss.

The design of this dynamic factor needs to satisfy two conditions: 1) it must be adjusted according to the continuous updating of the model training and 2) when one loss changes, the other reacts in a timely and reasonable response. Therefore, we first measure the distance between the WMMSD loss and the CORAL loss, in which the procedure is described as follows:

$$d_A = 2 \times \left(1 - 2 \times \sqrt{L_{\text{coral}}/L_{\text{wmmsd}}}\right). \tag{23}$$

Then, the value is mapped between [0, 1] by using sigmoid, which is described as follows:

$$\alpha = f_{\text{sigmoid}}(d_A). \tag{24}$$

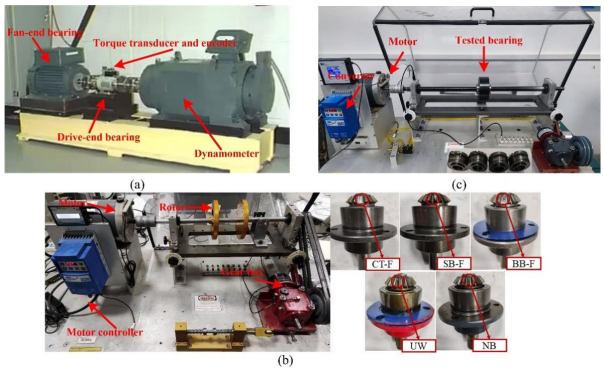


Fig. 5. Experimental equipment platform. (a) CWRU. (b) RG. (c) LZUT.

Finally, the JDM loss can be described as follows:

$$L_{\text{JDM}} = \alpha \cdot L_{\text{wmmsd}} + (1 - \alpha) \cdot L_{\text{coral}}. \tag{25}$$

4) Total Loss for the FJDMA: The total loss function of the proposed method consists of three parts: classifier loss L_{c}, domain discriminator loss L_{d}, and JDM loss L_{JDM}, which is described as follows:

$$L_{\text{total}} = L_{\text{c}} + L_{\text{d}} + L_{\text{JDM}}. \tag{26}$$

The flowchart of the pseudocode describing the training of the FJDMA is shown in Algorithm 1.

IV. Experimental Results and Discussion

In order to validate the fault diagnosis performance of the proposed method in cross-domain scenarios, we use Case Western Reserve University (CWRU) dataset, the Rolling Gear Simulation Test Bed (RG) dataset, and our own dataset (LZUT) for case studies. Specifically, we use CWRU for rolling bearing fault diagnosis in cross-domain scenarios and use RG for bevel gear fault diagnosis in cross-domain scenarios. In addition, we also perform the validation of the transfer learning performance between different datasets. The experimental environment is tensorflow 2.0.0-python 3.6.

A. Description of the Datasets

1) CWRU: The CWRU test platform is shown in Fig. 5(a). The tested objects of the dataset consisted of normal state, inner ring failure, outer ring failure, and rolling failure. Three sizes of damage 0.18, 0.36, and 0.54 mm are designed for the three failure states. In addition, the dataset collects the data under four operating conditions (0, 1, 2, and 3 HP). In this section, three fault states under three sizes and the normal state are selected to construct ten fault types for cross-domain fault diagnosis. The detailed information of the test data is shown in Table II. The 2048 sample points are taken as test data for each sample, where the training is 140 samples and the testing is 60 samples for each fault.

2) RG: RG is collected by ourselves and the test rig is shown in Fig. 5(b). The data are acquired by the shear acceleration sensor that the sequence number is LW65499. The bevel gear faults include small-end broken half-tooth failures (SB-F), big-end broken half-tooth failures (BB-F), uniform wear failures (UW), complete tooth breakage failures (CT-F), and normal bevel gear (NB). The data are obtained

TABLE II
DESCRIPTION OF THE DATASET

Dataset	Task	Conditions	Source/Training samples		Target/Test samples	
CWRU	A_0 A_1 A_2 A_3	1790rmp / 0HP 1772rmp / 1HP 1750rmp / 2HP 1730rmp / 3HP	N, 07B,14B,21B 07IR,14IR,21IR 07OR,14OR,21OR	1500	N, 07B,14B,21B 07IR,14IR,21IR 07OR,14OR,21OR	500
RG	B_0 B_1 B_2 B_3	0HP 1HP 2HP 3HP	CT-F, SB-F BB-F, UW NB	750	CT-F, SB-F BB-F, UW NB	250
LZUT	C_0 C_1 C_2 C_3	1449r/min (0HP) 1378r/min (1HP) 1251r/min (2HP) 1130r/min (3HP)	Ball, Inner Outer, Comb	600	Ball, Inner Outer, Comb	200

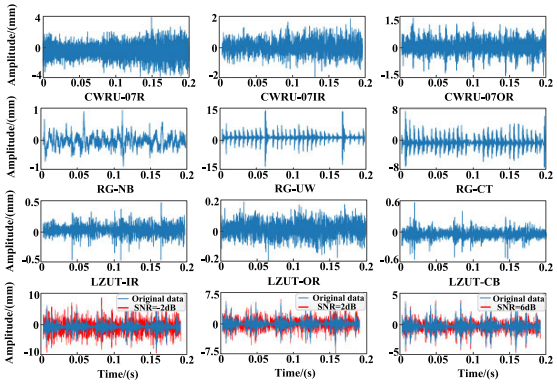


Fig. 6. Visualization results of vibration signals for different datasets.

for 0, 1, 2, and 3 HP at 20, 30, and 40 Hz by adjusting the inverter controller. In this section, four working conditions are selected for experimentation at 30 Hz. Table II shows the details of the dataset.

3) LZUT: LZUT is collected by ourselves and the test rig is shown in Fig. 5(c). The object to be tested is deep groove ball bearing of ER-16k in this test rig. The failure states of the bearing are rolling fault, inner ring fault, outer ring fault, combined inner, and outer ring fault. We collect the vibration signals under the rotational speeds of 1130 r/min (3 HP), 1251 r/min (2 HP), 1378 r/min (1 HP), and 1449 r/min (0 HP). The sampling frequency is 15.3 kHz and the collecting time is 8 s. Table II shows the details of the dataset. Fig. 6 shows the visualization of the collected signals and added noise from the experimental platform. As can be seen from Fig. 6, there is a significant difference in the collected vibration signals.

The network parameters are updated by the Adam optimizer and the learning rate is 0.001. The batch size is 64 and the epoch is 200.

B. Comparison Methods

To illustrate the superiority of the proposed method, the following DA methods are selected for comparison.

1) *DANN [32]:* It consists of feature extractor, predictive classifier, and domain discriminator. The model learns domain-invariant features by adversarial training.
2) *Deep Discriminative Transfer Learning Network (DDTLN) [33]:* It constructs a joint distribution adaptive module. The module consists of maximum mean discrepancy and CORAL.
3) *DADDN [20]:* It consists of dual domain adversarial mechanism and weighted joint adaptive distribution. The model uses balanced central weighting to reduce negative transfer.
4) *Joint Aiscriminative Adversarial Domain Adaptation (JAADA) [25]:* It is a joint-attention adversarial model. This model fuses local and global attention into DA to improve feature transferability.
5) *Joint Attention Adversarial Domain Adaptation (JDADA) [21]:* It achieves domain and class alignment by introducing class alignment into DA.
6) *Wavelet Information Domain Adaptation Network (WIDAN) [34]:* It is a physically informed wavelet domain adaptive network. This network integrates wavelet knowledge into convolutional layers for cross-domain machine fault diagnosis tasks.
7) *Physical Deep Spectral Network (PyDSN) [35]:* It is used for cross-domain fault diagnosis under speed fluctuations. It utilizes the modulated differentiable short-time Fourier transform to capture fault frequency information under speed fluctuations.
8) *Sparse Filtering Cross-Domain Adaptation (SFCDA) [3]:* It is a sparse filtered domain adaptive network. It uses soft reconstruction penalties to constrain the sparse filtering weights, thus improving feature independence.

C. Performance Comparison in Noisy Environment

In order to verify that the designed JDM is effective in reducing the two domain discrepancies and accelerating the network convergence. Fig. 7 plots the loss profile of the FJDMA. In Fig. 7, the classification loss is represented as the classifier loss value. From the figure, it can be seen that the WMMSD loss value and CORAL loss value fluctuate at the beginning of model training. As the number of epochs increases, the two loss values reach full convergence. In addition, the JDM loss values have the same trend. This observation suggests that the JDM loss can further reduce the two domain distribution differences.

In the early stage of rotating machinery, the fault features are not obvious due to the interference of noise. Therefore, the

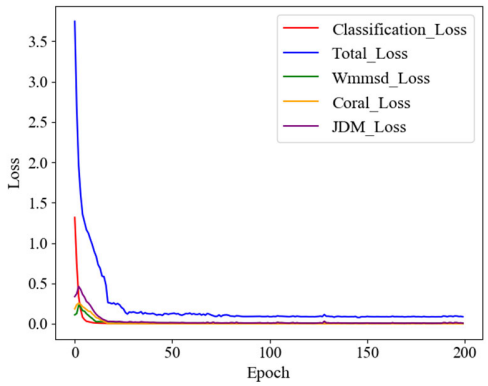


Fig. 7. Loss profile of the FJDMA.

TABLE III
DIAGNOSTIC ACCURACIES OF CWRU

(%)	-2dB	0dB	2dB	4dB	6dB	W-snr
A_0-A_1	84.46	90.86	98.66	99.33	99.93	100.0
A_0-A_2	85.30	91.46	98.13	99.13	99.73	100.0
A_0-A_3	89.73	95.13	98.60	99.73	99.93	100.0
A_1-A_0	82.06	93.60	97.93	98.93	99.40	99.93
A_1-A_2	87.47	94.00	98.40	99.66	99.80	100.0
A_1-A_3	88.06	93.73	96.06	97.86	99.80	99.97
A_2-A_0	86.47	94.80	96.66	98.26	99.20	99.91
A_2-A_1	85.80	95.80	96.06	99.66	99.93	100.0
A_2-A_3	84.80	93.81	97.86	99.46	99.97	100.0

fault diagnosis model should have stable and reliable resistance to noise interference. In this section, we use the signal-to-noise ratio (SNR) to measure the strength of vibration signal and noise, which is defined as follows:

$$X_{\text{snr}} = 20\lg(A_{\text{signal}}/A_{\text{noise}}) \tag{27}$$

where A_{signal} is the vibration signal amplitude and A_{noise} is the noise amplitude. Fig. 6 shows the visualization of the different datasets after adding noise. It is worth noting that we only add noise to the target domain in the noise experiment to better validate the transfer task in an unknown environment.

Fig. 8 and Table III show the transfer diagnosis results of CWRU in different noise environments. As can be seen from Fig. 8, when SNR $= -2$ dB, the diagnostic accuracy of FJDMA is the lowest in different transfer tasks. This is because we only add noise signals to the target domain and do not add noise signals to the source domain. In the strong noise environment, since the noise signal has a large interference to the vibration signal, it makes the fault features more complicated leading to a decrease in the diagnostic accuracy for the model.

When adding 6-dB noise, the diagnostic accuracy of FJDMA is above 99.0%. This indicates that with the weakening of the noise interference, the ability of the model to capture fault features is increasing. Table III shows that the overall diagnostic accuracy of the model is above 80% when we add SNR $= -2$ dB to SNR $= 6$ dB to the CWRU dataset. This indicates that FJDMA has noise resistance and is able to satisfy transfer learning fault diagnosis in strong noise environments.

Fig. 9 and Table IV show the transfer diagnosis results of RG in noisy environments. From Fig. 9, we see that the accuracy of FJDMA is above 80% under different noise environments. This indicates that FJDMA can meet the diagnosis requirements under noise environments. The FJDMA has an average accuracy of 88.55% under nine different tasks at SNR $= -4$ dB. This indicates that FJDMA has good noise immunity. In addition, the error line in Fig. 9 also shows that FJDMA also has good stability. The average accuracy of FJDMA is 99.74% in nine tasks at SNR $= 4$ dB. This indicates that FJDMA has robustness and can satisfy the fault diagnosis tasks in noisy environments.

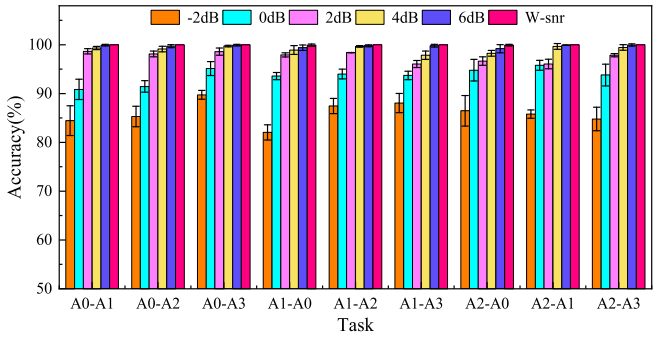


Fig. 8. Performance comparison of CWRU under noise environments.

TABLE IV
DIAGNOSTIC ACCURACIES OF RG

(%)	-4dB	-2dB	0dB	2dB	4dB	W-snr
B_0-B_1	83.33	94.26	96.53	98.53	99.73	100.0
B_0-B_2	83.54	94.00	97.20	98.26	99.60	100.0
B_0-B_3	86.40	94.13	96.93	97.33	99.73	100.0
B_1-B_0	82.53	90.80	96.13	97.06	99.73	99.97
B_1-B_2	82.13	92.00	97.06	98.66	100.0	100.0
B_1-B_3	81.73	92.00	97.60	99.33	99.86	99.98
B_2-B_0	88.26	94.13	98.53	99.46	99.93	100.0
B_2-B_1	84.80	94.00	96.20	99.33	99.86	99.96
B_2-B_3	88.26	90.00	97.60	99.06	99.80	100.0

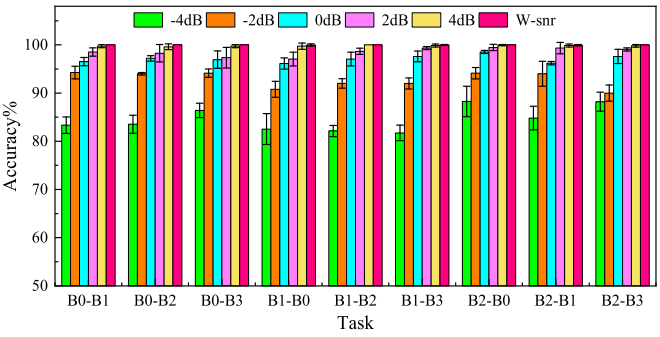


Fig. 9. Performance comparison of RG under noise environments.

D. Cross-Domain Diagnostic Results Under Variable Loads

To visually evaluate the performance of FJDMA and compare the feature distributions in the source and target domains, we use the t-SNE technique to visualize and plot the probability density distributions (PDFs). Specifically, we select A_0–A_3 for visualization in the CWRU dataset. In addition, to better simulate the unknown working conditions, we only add -2-dB noise to the target domain, and the visualization results are shown in Fig. 10. From the pdf in Fig. 10(a), we see that the PDFs of the source and target domains match poorly, which indicates that the fault features have obvious domain differences under different operating conditions. DANN cannot

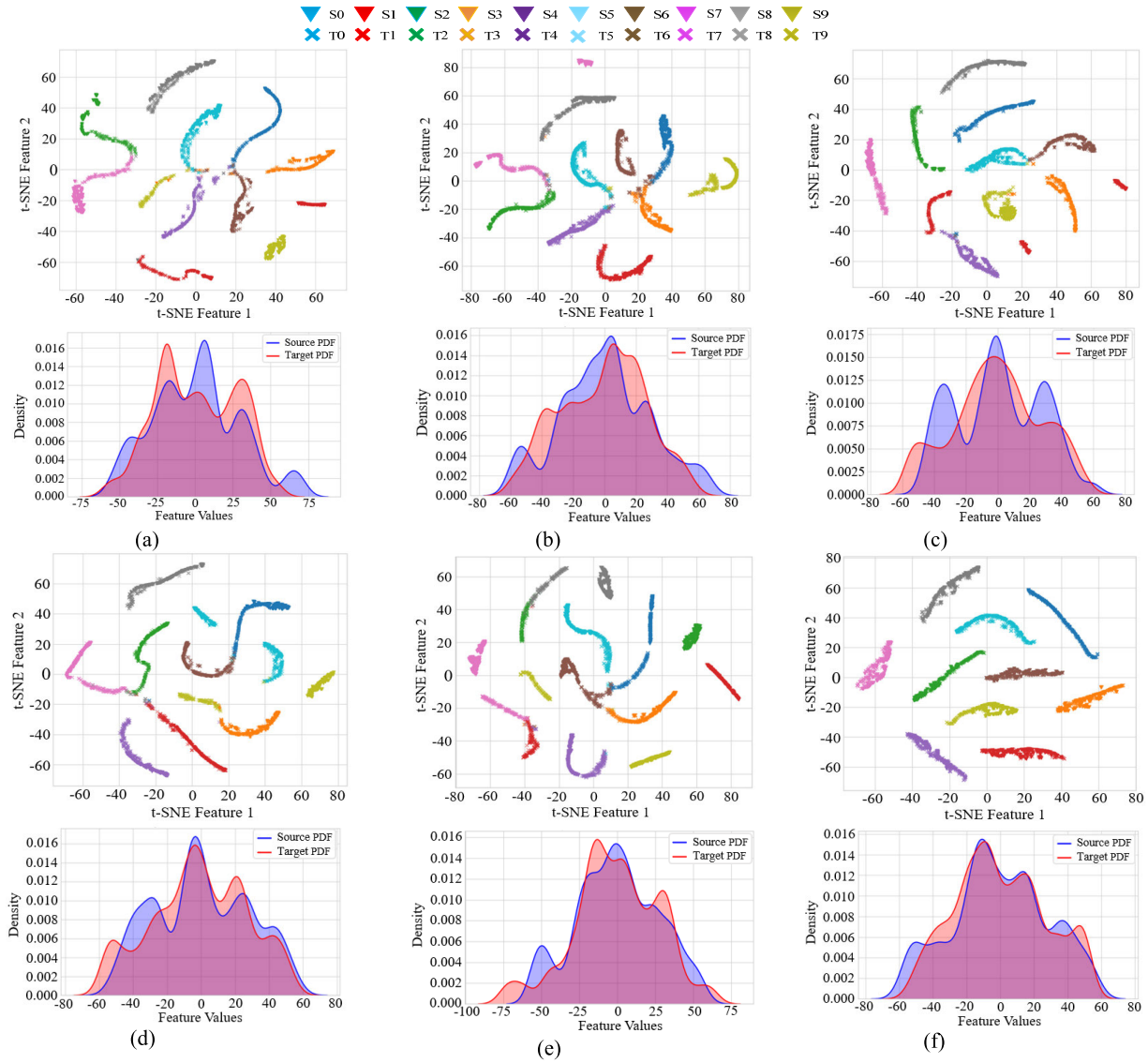


Fig. 10. Visualization results of CWRU. (a) DANN. (b) DDTLN. (c) DADDN. (d) JAADA. (e) JDADA. (f) Proposed.

effectively classify cross-domain faults, which can also be seen in the visualization of t-SNE.

However, when we use some domain adaptive techniques, the overlapping part of source and target domain pdf increases compared to DANN. At the same time, we can clearly observe from the visualization of the 2-D features that the classification of different classes is significantly improved. In Fig. 10(b), DDTLN focuses on the interdomain differences between categories using coral distribution. However, when using MMD to reduce the intradomain differences, it is difficult to identify the same fault type because MMD uses the mean as the distribution distance, for example, source and target domain labels 3 (07OR) and 6 (14OR). In Fig. 10(c), DADDN balanced central weighting can focus on the differences between classes in the interdomain, resulting in increased distances between classes. However, the method does fail to reduce the intraclass distance, affecting classification accuracy. In Fig. 10(d), JAADA uses the attention mechanism to focus more on useful features, resulting in a further reduction in intraclass distance. However, the method is limited in its ability to recognize interdomain class differences and cannot effectively increase the interclass distance. In Fig. 10(f), different classes are separated and have large interclass distances, the same classes are aggregated together and have smaller intraclass distances. This indicates that FJDMA can pay more attention to useful features and to mine transferable features.

We select B_0–B_3 in the RG dataset and add SNR $= -4$ dB to the target domain, and the visualization results are shown in Fig. 11. In Fig. 11, the pdf curves of the comparison methods are poorly matched, which indicates that the comparison methods have a lack of ability to capture the differences between fault features. However, the source and target domain PDFs of the proposed FJDMA mostly overlap. Specifically, label 3 (complete tooth breaking faults) and label 4 (large-end breaking half-tooth faults) are mixed in Fig. 11(a)–(c) and cannot be classified effectively. This is because the fault features of labels 3 and 4 are similar, and the limited feature extraction capability of DANN, DDTLN, and DADDN makes it difficult to identify their faults. In Fig. 11(d) and (e), the interdomain class distances are increased, which indicates that JAADA and JDADA pay more attention to the domain

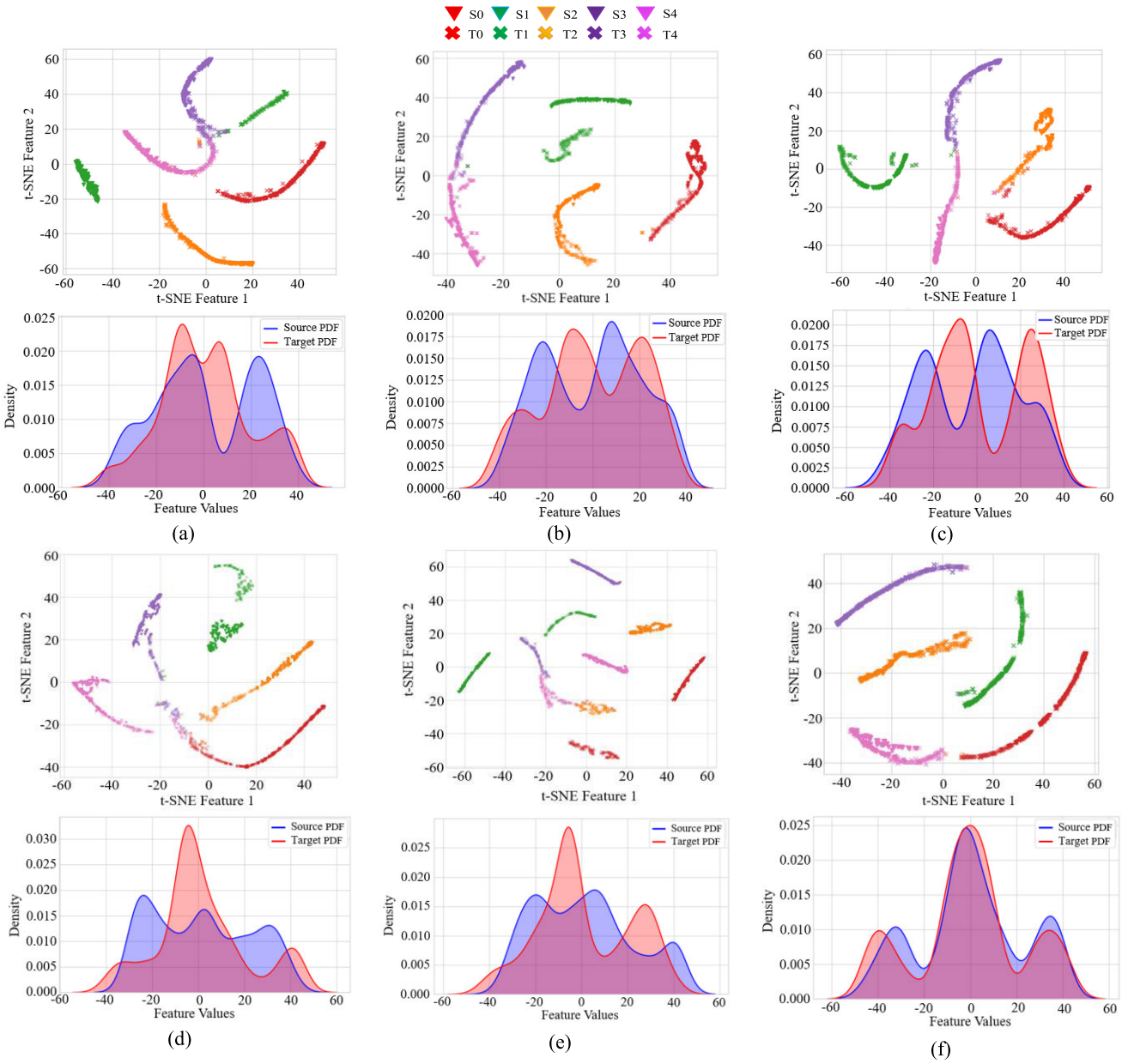


Fig. 11. Visualization results of RG. (a) DANN. (b) DDTLN. (c) DADDN. (d) JAADA. (e) JDADA. (f) Proposed.

feature differences. However, JAADA and JDADA cannot reduce the intraclass distance, and the source and target domain features cannot be clustered together, which affects the transfer learning performance. In Fig. 11(f), different classes are separated and have large interclass distances. The same classes are aggregated together. This indicates that FJDMA can focus more on useful features and has better classification performance.

E. Cross-Domain Diagnostic Results Between Different Datasets

To rigorously test the experimental conditions and further verify the generalization performance of FJDMA, we conduct transfer learning experiments between CWRU, RG, and LZUT.

Among them, the results of CWRU and RG experiments are shown in Table V and Fig. 12. In Table V, the fault diagnosis accuracies of the comparison methods are lower than that of FJDMA in all eight migration tasks. Specifically, the average diagnostic accuracy of FJDMA is 11.02% higher than DANN in all transfer tasks, which indicates that the designed feature alignment module and the joint distributional migration can capture the transferable features. The average diagnostic accuracy of FJDMA is improved by 7.18%, 6.51%, and 12.99% compared to DDTLN, DADDN, and JDADA in all transfer tasks, respectively. This indicates that only focusing on interdomain differences is insufficient in cross-domain fault diagnosis. The average diagnostic accuracy of FJDMA is improved by 6.56% compared to JAADA, which indicates that the joint distributional migration improves the transfer learning performance of the model in cross-domain. In addition, the average diagnostic accuracy of the FJDMA is improved by 2.38% and 4.34% compared to WIDAN and PyDSN, respectively. This suggests that transferable features can be enhanced by introducing physical information knowledge. However, WIDAN and PyDSN are not aligned with features through domain adversarial training,

TABLE V
CROSS-DOMAIN DIAGNOSTIC RESULT FOR CWRU AND RG

Task	DANN	DDTLN	DADDN	JAADA	JDADA	WIDAN	PyDSN	SFCDA	Proposed
A_0-B_0	81.90±3.91	85.20±1.04	84.40±2.24	85.46±3.06	74.80±2.84	88.80±1.34	86.76±0.14	85.13±0.32	92.08±2.72
A_1-B_1	79.60±1.08	79.86±2.88	84.80±0.68	86.26±0.67	78.32±1.66	84.26±0.19	82.41±0.33	82.06±0.24	88.04±1.96
A_2-B_2	78.36±0.80	84.00±0.64	84.66±1.73	85.60±0.32	76.34±0.73	85.66±0.13	85.02±0.10	83.19±0.49	86.52±0.92
A_3-B_3	80.44±1.26	83.06±3.86	87.46±0.46	84.13±2.13	84.40±3.06	88.02±0.03	85.88±0.54	86.01±0.03	89.92±0.48
B_0-A_0	80.40±1.16	80.93±0.93	84.53±1.06	87.20±1.12	72.48±1.21	86.31±0.76	83.33±1.17	82.76±0.52	89.60±2.80
B_1-A_1	82.82±1.60	86.13±0.67	85.80±0.76	85.42±1.46	84.64±1.44	91.69±1.02	88.24±0.55	85.38±0.07	94.04±0.80
B_2-A_2	54.44±2.40	64.13±0.13	59.43±0.63	59.20±0.80	57.60±0.23	77.08±0.02	74.32±0.60	69.95±0.36	77.92±0.32
B_3-A_3	81.60±2.06	86.93±1.06	84.53±2.13	82.00±0.80	75.23±2.36	86.88±0.84	87.04±0.27	84.92±0.36	89.61±1.41

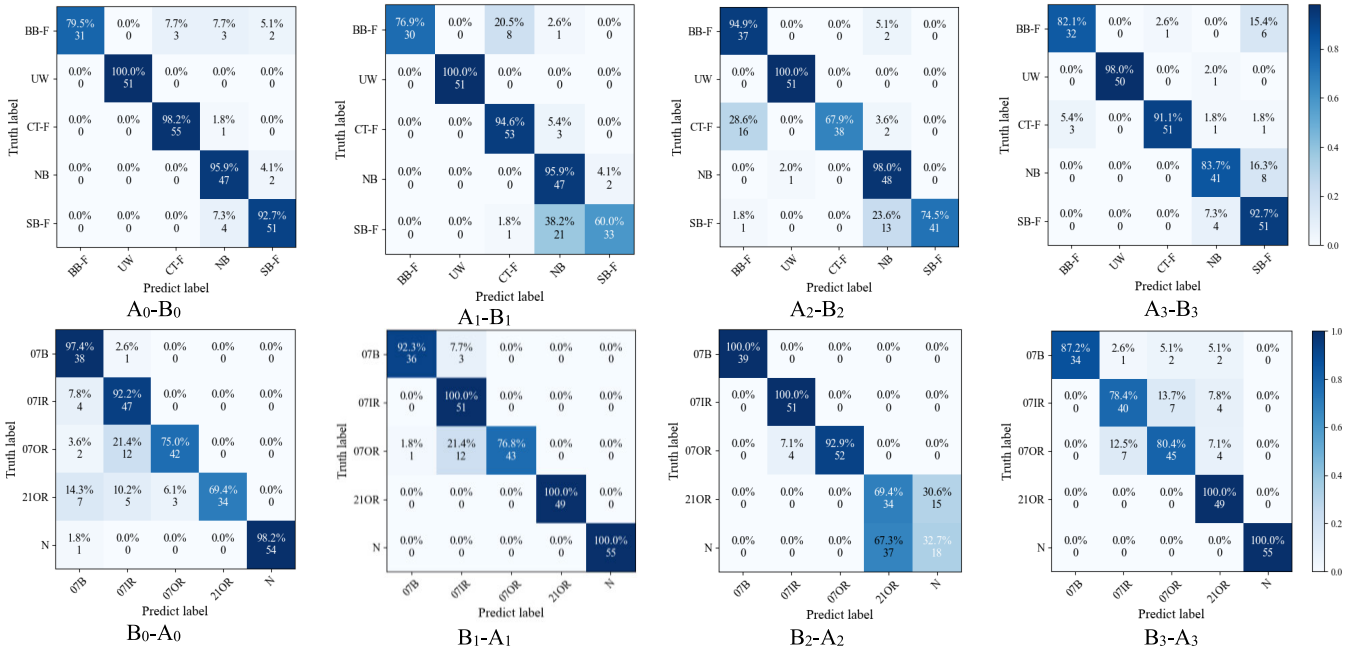


Fig. 12. Diagnostic results of FJDMA between CWRU and RG datasets.

which leads to a decrease in the diagnostic accuracy of the models. The average diagnostic accuracy of the proposed method is improved by 6.03% compared with SFCDA. This indicates that the FJDMA performs domain alignment in terms of features and metrics, which can effectively reduce the domain discrepancy and enhance diagnostic accuracy.

Fig. 12 shows the confusion matrix results of FJDMA between CWRU and RG. In Fig. 12, FJDMA has a high misclassification rate for BB-F and SB-F in the transfer task of A–B. This indicates that the fault features of BB-F and SB-F are more similar, which causes FJDMA to easily misclassify them as other faults during fault identification. However, FJDMA has an average accuracy of 89.14% in the four transfer tasks of A–B, which are all higher than the comparison methods. In the transfer task of B–A, FJDMA has a higher misclassification rate for 07OR and 21OR, which indicates that FJDMA learns insufficient domain-invariant features when the RG dataset is used as the source domain due to the large data difference between the source and target domains. However, FJDMA has an average accuracy of 87.79% in the four transfer tasks, which is higher than the comparison methods.

Table VI shows the transfer results on different test platforms where the tested objects are all bearings. From Table VI, the fault identification accuracy of FJDMA is above 90% in all eight transfer tasks, and its average diagnosis accuracy is 95.01%. This further indicates that FJDMA has good cross-domain diagnosis performance. Specifically, the average diagnosis accuracy of FJDMA is improved by 6.03%, 6.37%, and 8.99% compared to DDTLN, DADDN, and JDADA, respectively. This indicates that it is necessary to enhance the transferable features and suppress the redundant features through the attention mechanism. The average diagnostic accuracy of FJDMA is improved by 8.77% compared to JAADA, which indicates that the designed joint distributional migration can effectively identify domain differences and improve the diagnostic accuracy. In addition, the average diagnostic accuracy of the FJDMA is improved by 3.41% and 5.00% compared to the state-of-the-art WIDAN and PyDSN, respectively. The average diagnostic accuracy of the FJDMA is improved by 8.16% compared to SFCDA. Although WIDAN and PyDSN can enhance the domain adaptive capability by including physical information in the model. However, the model does not perform domain adversarial learning resulting in poor diagnostic accuracy. The SFCDA only uses two layers of FC as a feature extractor, resulting in insufficient feature extraction capability of the model.

TABLE VI
CROSS-DOMAIN DIAGNOSTIC RESULT FOR CWRU AND LZUT

Task	DANN	DDTLN	DADDN	JAADA	JDADA	WIDAN	PyDSN	SFCDA	Proposed
A_0-C_0	83.16±2.33	86.44±1.91	87.00±0.96	81.75±0.76	79.16±1.83	88.30±1.37	89.68±0.14	86.48±0.37	91.66±1.16
A_1-C_1	86.66±1.31	90.05±0.23	90.50±1.58	84.54±1.12	86.05±0.32	90.12±0.59	89.00±0.76	87.92±0.12	95.00±0.17
A_2-C_2	84.12±1.80	87.24±1.36	87.06±0.33	84.24±0.66	85.50±1.33	89.80±0.23	87.02±0.22	85.00±0.61	93.50±0.50
A_3-C_3	84.05±0.94	91.55±0.16	88.50±2.66	87.05±1.06	90.04±0.16	94.34±0.60	92.13±1.44	89.65±0.17	96.16±1.33
C_0-A_0	85.46±0.16	90.93±1.03	90.14±1.96	88.50±2.24	86.64±1.72	93.64±0.07	90.97±0.55	87.08±0.63	96.71±0.32
C_1-A_1	85.00±2.64	86.13±0.67	91.33±0.44	90.50±0.46	88.34±0.84	93.66±0.56	90.40±1.08	87.43±0.75	95.12±0.56
C_2-A_2	85.04±1.42	88.00±1.73	88.54±0.22	87.34±0.88	85.25±1.56	92.97±0.03	89.86±0.17	85.90±0.54	96.66±0.16
C_3-A_3	85.76±1.31	91.50±0.40	86.05±1.76	86.00±0.12	87.24±0.98	90.02±0.81	91.06±0.66	85.32±0.80	95.33±0.45

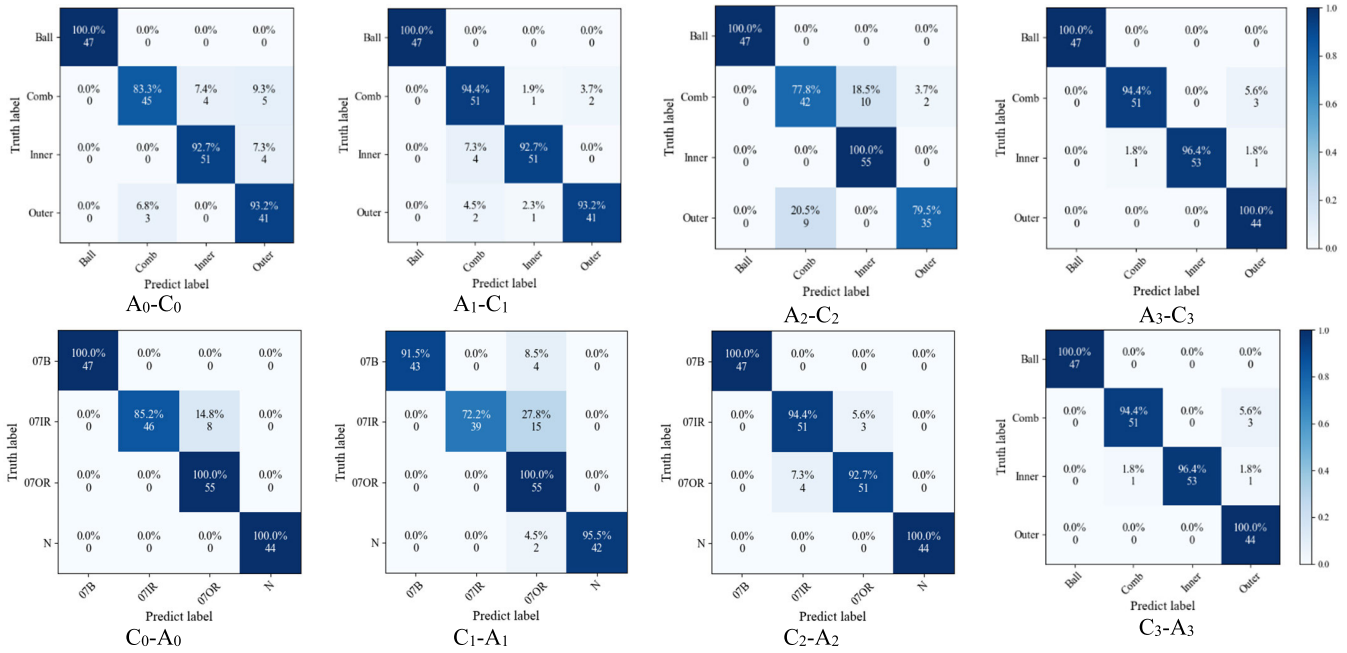
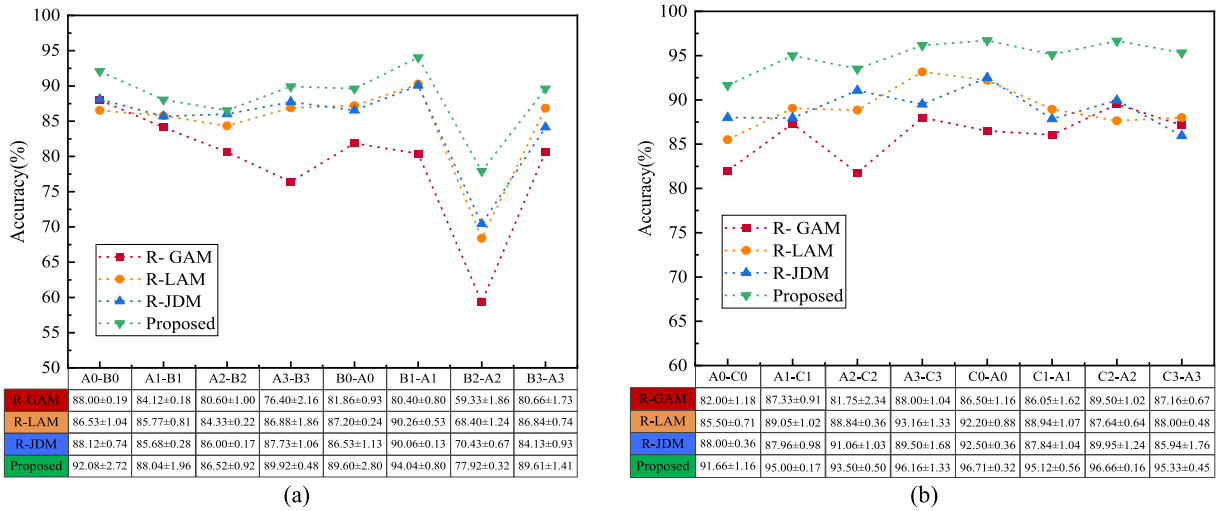


Fig. 13. Diagnostic results of FJDMA between CWRU and LZUT datasets.



	A0-B0	A1-B1	A2-B2	A3-B3	B0-A0	B1-A1	B2-A2	B3-A3
R-GAM	88.00±0.19	84.12±0.18	80.60±1.00	76.40±2.16	81.86±0.93	80.40±0.80	59.33±1.86	80.66±1.73
R-LAM	86.53±1.04	85.77±0.81	84.33±0.22	86.88±1.86	87.20±0.24	90.26±0.53	68.40±1.24	86.84±0.74
R-JDM	88.12±0.74	85.68±0.28	86.00±0.17	87.73±1.06	86.53±1.13	90.06±0.13	70.43±0.67	84.13±0.93
Proposed	92.08±2.72	88.04±1.96	86.52±0.92	89.92±0.48	89.60±2.80	94.04±0.80	77.92±0.32	89.61±1.41

(a)

	A0-C0	A1-C1	A2-C2	A3-C3	C0-A0	C1-A1	C2-A2	C3-A3
R-GAM	82.00±1.18	87.33±0.91	81.75±2.34	88.00±1.04	86.50±1.16	86.05±1.62	89.50±1.02	87.16±0.67
R-LAM	85.50±0.71	89.05±1.02	88.84±0.36	93.16±1.33	92.20±0.88	88.94±1.07	87.64±0.64	88.00±0.48
R-JDM	88.00±0.36	87.96±0.98	91.06±1.03	89.50±1.68	92.50±0.36	87.84±1.04	89.95±1.24	85.94±1.76
Proposed	91.66±1.16	95.00±0.17	93.50±0.50	96.16±1.33	96.71±0.32	95.12±0.56	96.66±0.16	95.33±0.45

(b)

Fig. 14. Diagnostic results of ablation experiments. (a) Cross-domain diagnostic results for CWRU and RG. (b) Cross-domain diagnostic results for CWRU and LZUT.

Fig. 13 shows the confusion matrix results of FJDMA for cross-domain diagnosis between CWRU and LZUT. From Fig. 12, we see that FJDMA has a higher misclassification rate for Comb (combined inner and outer faults) in the A–C transfer task, this is because this fault type is more complex, and FJDMA is prone to misclassify it as an inner or outer ring fault. However, the fault identification accuracy for other fault types is higher and its average diagnostic accuracy is 95.01%. This indicates that FJDMA can mine domain-invariant features and align two domains at the feature aspect, which improves the fault recognition accuracy under cross-domain.

F. Ablation Experiments

In this section, in order to verify the impact of each component on the model performance, we conduct ablation experiments. The specific experimental setup is as follows: removing the local alignment module (R-LAM), removing the global alignment module (R-GAM), and removing the joint distributional migration (R-JDM). The experimental results are shown in Fig. 14. In Fig. 14, R-GAM has the greatest impact on the diagnostic performance of the model. Its average diagnostic accuracy is 9.54% and 8.98% lower than the proposed methods. This indicates that the global alignment module can pay more attention to the transferable features and significantly improve the transfer effect of the model. When there is no global alignment module, the diagnostic accuracy of the proposed method decreases significantly. In Fig. 14(a), the average diagnostic accuracy of the proposed method is improved by 3.63% and 3.93% over R-JDM and R-LAM, respectively. This indicates that the JDM can reduce the two-domain distribution difference from the metric distance aspect during model training, and the local alignment module can over further focus on more domain-invariant features from the feature aspect. In addition, the average diagnostic accuracy of FJDMA is 95.02% in the dataset of bearings when all modules are activated. The average diagnostic accuracy is 88.47% in the dataset where the tested objects are bearings and gears.

This further illustrates that the proposed method has superior performance and better generalization performance for cross-domain fault diagnosis. In Fig. 14(a), we also notice that the diagnostic accuracy of the proposed method shows a significant decrease in the B_2–A_2 migration task. This is because the source domain is the gear dataset and the target domain is the bearing dataset in the B_2–A_2 migration task. Thus, there are large differences in the samples of the source and target domains, and the big-end broken half-tooth faults have similar fault characteristics to the outer ring faults, which leads to the insufficient domain-invariant features learned by the proposed method.

To further visualize the contribution of each module to feature metastability, we selected the second feature extraction module of the feature extractor for feature visualization, as shown in Fig. 15. In Fig. 15, the horizontal axis is the number of channels and the vertical axis is the time index. Different colors represent different feature values. In Fig. 15(a), the higher feature values are shown to be richer and more homogeneous, which indicates that the proposed method can focus on more domain-invariant features. In Fig. 15(b) and (c), we find that the distribution of higher feature values starts to decrease. This suggests that R-LAM and the JDM reduce the model's ability to capture domain-invariant features, leading to a decrease in the model's diagnostic performance. In Fig. 15(d), we find that most of the feature values are low and the distribution of higher feature values is rare. This indicates that the global alignment module has the greatest impact on the model's performance. The diagnostic performance of the model is severely degraded after R-GAM.

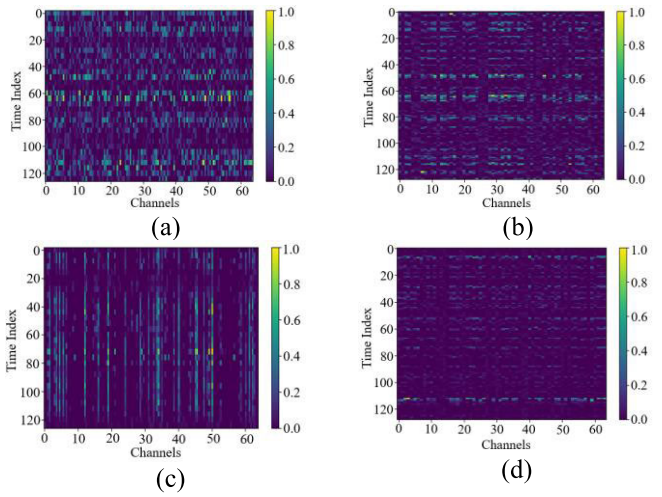


Fig. 15. Results of feature visualization. (a) Proposed. (b) R-LAM. (c) R-JDM. (d) R-GAM.

V. Conclusion

In this article, we propose a domain adversarial fault diagnosis method based on FJDMA. The method performs domain adaptation at the feature and domain training aspects. Specifically, the following conditions hold: 1) a feature aligner is designed at the feature aspect, which learns domain-invariant features from both local and global perspectives; and 2) a joint distributional migration is designed at the domain training aspect, which consists of the weighted maximum mean square discrepancy loss and the CORAL loss to achieve intradomain and interdomain classes distribution alignment. Finally, experimental analyses are performed by two bearing datasets and one gearbox dataset and compared with state-of-the-art methods. The results show that FJDMA has the highest diagnostic accuracy in both noisy environments and cross-domain fault diagnosis.

In this article, we study single-source domain cross-domain fault diagnosis. In the future, we will extend FJDMA to multisource domain scenarios.

References

[1] Y. Zhang, Z. Ren, K. Feng, K. Yu, M. Beer, and Z. Liu, "Universal source-free domain adaptation method for cross-domain fault diagnosis of machines," *Mech. Syst. Signal Process.*, vol. 191, May 2023, Art. no. 110159.

[2] L. Zhang, H. Zhang, and G. Cai, "The multiclass fault diagnosis of wind turbine bearing based on multisource signal fusion and deep learning generative model," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[3] F. Lu et al., "Towards multi-scene learning: A novel cross-domain adaptation model based on sparse filter for traction motor bearing fault diagnosis in high-speed EMU," *Adv. Eng. Informat.*, vol. 60, Apr. 2024, Art. no. 102536.

[4] Y. Xiao, H. Shao, J. Wang, S. Yan, and B. Liu, "Bayesian variational transformer: A generalizable model for rotating machinery fault diagnosis," *Mech. Syst. Signal Process.*, vol. 207, Jan. 2024, Art. no. 110936.

[5] Y. Xiao, H. Shao, M. Feng, T. Han, J. Wan, and B. Liu, "Towards trustworthy rotating machinery fault diagnosis via attention uncertainty in transformer," *J. Manuf. Syst.*, vol. 70, pp. 186–201, Oct. 2023.

[6] J. Peng, H. Shao, Y. Xiao, B. Cai, and B. Liu, "Industrial surface defect detection and localization using multi-scale information focusing and enhancement GANomaly," *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 122361.

[7] J. Lin, H. Shao, X. Zhou, B. Cai, and B. Liu, "Generalized MAML for few-shot cross-domain fault diagnosis of bearing driven by heterogeneous signals," *Expert Syst. Appl.*, vol. 230, Nov. 2023, Art. no. 120696.

[8] W. Ma, Y. Zhang, L. Ma, R. Liu, and S. Yan, "An unsupervised domain adaptation approach with enhanced transferability and discriminability for bearing fault diagnosis under few-shot samples," *Expert Syst. Appl.*, vol. 225, Sep. 2023, Art. no. 120084.

[9] C. Zhao and W. Shen, "Dual adversarial network for cross-domain open set fault diagnosis," *Rel. Eng. Syst. Saf.*, vol. 221, May 2022, Art. no. 108358.

[10] Z. Wu, G. Fang, Y. Wang, and R. Xu, "An end-to-end deep clustering method with consistency and complementarity attention mechanism for multisensor fault diagnosis," *Appl. Soft Comput.*, vol. 158, Jun. 2024, Art. no. 111594.

[11] X. Yang, X. Yuan, T. Ye, W. Zhu, F. Zhou, and J. Jin, "PSNN-TADA: Prototype and stochastic neural network-based twice adversarial domain adaptation for fault diagnosis under varying working conditions," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–12, 2024.

[12] H. Shao, M. Xia, G. Han, Y. Zhang, and J. Wan, "Intelligent fault diagnosis of rotor-bearing system under varying working conditions with modified transfer convolutional neural network and thermal images," *IEEE Trans. Ind. Informat.*, vol. 17, no. 5, pp. 3488–3496, May 2021.

[13] F. Liu et al., "Structural discrepancy and domain adversarial fusion network for cross-domain fault diagnosis," *Adv. Eng. Informat.*, vol. 58, Oct. 2023, Art. no. 102217.

[14] S. Yan, H. Shao, J. Wang, X. Zheng, and B. Liu, "LiConvFormer: A lightweight fault diagnosis framework using separable multiscale convolution and broadcast self-attention," *Expert Syst. Appl.*, vol. 237, Mar. 2024, Art. no. 121338.

[15] X. Zhao and Y. Zhang, "An intelligent diagnosis method of rolling bearing based on multi-scale residual shrinkage convolutional neural network," *Meas. Sci. Technol.*, vol. 33, no. 8, Aug. 2022, Art. no. 085103.

[16] F. Jia, Y. Wang, J. Shen, L. Hao, and Z. Jiang, "Stepwise feature norm network with adaptive weighting for open set cross-domain intelligent fault diagnosis of bearings," *Meas. Sci. Technol.*, vol. 35, no. 5, May 2024, Art. no. 056126.

[17] C. He, H. Shi, and J. Li, "IDSN: A one-stage interpretable and differentiable STFT domain adaptation network for traction motor of high-speed trains cross-machine diagnosis," *Mech. Syst. Signal Process.*, vol. 205, Dec. 2023, Art. no. 110846.

[18] Y. Xiao, H. Shao, Z. Min, H. Cao, X. Chen, and J. Lin, "Multiscale dilated convolutional subdomain adaptation network with attention for unsupervised fault diagnosis of rotating machinery cross operating conditions," *Measurement*, vol. 204, Nov. 2022, Art. no. 112146.

[19] T. Han et al., "Novel adaptive loss weighted transfer network for partial domain fault diagnosis," *ISA Trans.*, vol. 145, pp. 362–372, Feb. 2024.

[20] F. Liu, W. Deng, C. Duan, Y. Qin, J. Luo, and H. Pu, "Duplex adversarial domain discriminative network for cross-domain partial transfer fault diagnosis," *Knowl.-Based Syst.*, vol. 279, Nov. 2023, Art. no. 110960.

[21] K. Sun, X. Xu, N. Lu, H. Xia, and M. Han, "Joint discriminative adversarial domain adaptation for cross-domain fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023.

[22] H. Liang, J. Cao, and X. Zhao, "Multibranch and multiscale dynamic convolutional network for small sample fault diagnosis of rotating machinery," *IEEE Sensors J.*, vol. 23, no. 8, pp. 8973–8988, Apr. 2023.

[23] X. Zhao and W. Luo, "A deep intelligent hybrid model for fault diagnosis of rolling bearing," *J. Vibrat. Eng. Technol.*, vol. 11, no. 2, pp. 721–737, Feb. 2023.

[24] H. Lv, J. Chen, T. Pan, T. Zhang, Y. Feng, and S. Liu, "Attention mechanism in intelligent fault diagnosis of machinery: A review of technique and application," *Measurement*, vol. 199, Aug. 2022, Art. no. 111594.

[25] P. Chen, R. Zhao, T. He, K. Wei, and J. Yuan, "A novel bearing fault diagnosis method based joint attention adversarial domain adaptation," *Rel. Eng. Syst. Saf.*, vol. 237, Sep. 2023, Art. no. 109345.

[26] X. Shao and C.-S. Kim, "Adaptive multi-scale attention convolution neural network for cross-domain fault diagnosis," *Expert Syst. Appl.*, vol. 236, Feb. 2024, Art. no. 121216.

[27] Y. Yao, Q. Chen, G. Gui, S. Yang, and S. Zhang, "A hierarchical adversarial multi-target domain adaptation for gear fault diagnosis under variable working condition based on raw acoustic signal," *Eng. Appl. Artif. Intell.*, vol. 123, Aug. 2023, Art. no. 106449.

[28] Y. Xiao, H. Shao, S. Han, Z. Huo, and J. Wan, "Novel joint transfer network for unsupervised bearing fault diagnosis from simulation domain to experimental domain," *IEEE/ASME Trans. Mechatronics*, vol. 27, no. 6, pp. 5254–5263, Dec. 2022.

[29] X. Yu, Y. Wang, Z. Liang, H. Shao, K. Yu, and W. Yu, "An adaptive domain adaptation method for rolling bearings' fault diagnosis fusing deep convolution and self-attention networks," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–14, 2023.

[30] J. Wang, H. Ren, C. Shen, W. Huang, and Z. Zhu, "Multi-scale style generative and adversarial contrastive networks for single domain generalization fault diagnosis," *Rel. Eng. Syst. Saf.*, vol. 243, Mar. 2024, Art. no. 109879.

[31] Q. Qian, Y. Wang, T. Zhang, and Y. Qin, "Maximum mean square discrepancy: A new discrepancy representation metric for mechanical fault transfer diagnosis," *Knowl.-Based Syst.*, vol. 276, Sep. 2023, Art. no. 110748.

[32] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1–35, Jan. 2016.

[33] Q. Qian, Y. Qin, J. Luo, Y. Wang, and F. Wu, "Deep discriminative transfer learning network for cross-machine fault diagnosis," *Mech. Syst. Signal Process.*, vol. 186, Mar. 2023, Art. no. 109884.

[34] C. He, H. Shi, X. Liu, and J. Li, "Interpretable physics-informed domain adaptation paradigm for cross-machine transfer diagnosis," *Knowledge-Based Syst.*, vol. 288, Mar. 2024, Art. no. 111499.

[35] C. He, H. Shi, R. Li, J. Li, and Z. Yu, "Interpretable modulated differentiable STFT and physics-informed balanced spectrum metric for freight train wheelset bearing cross-machine transfer fault diagnosis under speed fluctuations," *Adv. Eng. Informat.*, vol. 62, Oct. 2024, Art. no. 102568.

Measurement 251 (2025) 117278

Contents lists available at ScienceDirect

Measurement

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

Cross-domain remaining useful life prediction for rolling bearings based on wavelet decomposition and dynamic calibrated domain adaptive networks

Yazhou Zhang [a], Xiaoqiang Zhao [a,b,*], Zhenrui Peng [a,b], Rongrong Xu [a], Yongyong Hui [a,b]

[a] College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China
[b] Gansu Key Laboratory of Advanced Control of Industrial Processes, Lanzhou 730050, China

ARTICLE INFO

Keywords:
Rolling bearing
Remaining useful life prediction
Domain adaptive
Variable operating conditions
Dynamic temporal calibration unit
Broadcast multi-head self-attention gated recurrent unit

ABSTRACT

Domain Adaptive (DA) is widely used for the cross-domain remaining useful life (RUL) prediction of rolling bearings. However, most of the existing DA-based RUL methods reduce the distribution discrepancy from distance metric, ignoring the impact of degraded features on transfer learning. Therefore, this paper proposes a wavelet decomposition and dynamic calibration domain adaptive network (DCDAN). DCDAN enhances the domain representation in three dimensions: input samples, feature extraction and distance metric. Specifically, the data enhancement module is designed for the input samples, which performs multi-resolution analysis by wavelet transform to obtain different levels of low-frequency and high-frequency signals. Meanwhile, the frequency enhancement channel attention mechanism is used to preserve critical degradation features and reduce the missing of frequency information. The multi-scale temporal convolution module (MTCM) and the broadcast multi-head self-attention gated recurrent unit (BMSA-GRU) are designed for the shared feature extractor to extract more domain invariant features. In addition, when performing domain-adversarial training, a smooth multi-kernel maximum mean difference is designed to further reduce the difference between the source and target domains. Finally, the effectiveness of DCDAN is validated on two bearing datasets, the results show that DCDAN has better RUL prediction accuracy than other methods.

1. Introduction

Rolling bearing is an indispensable part of the rotating machinery, which is widely used in various industrial fields. Its health state determines the useful life of the whole rotating machinery [1,2]. Therefore, the remaining useful life (RUL) prediction of rolling bearings can help to repair mechanical equipment in time, and prolong the life of mechanical equipment, and avoid major safety accidents [3–5].

In recent years, scholars have proposed a lot of RUL prediction methods for rolling bearings and have achieved fruitful research results. These methods can be classified into two categories: model-driven based methods and data-driven based methods [6–8]. Model-driven based methods require that the mathematical model be constructed to describe the degradation process for rolling bearings. However, the complex and changing working environment makes it difficult to build an accurate mathematical model and describe the degradation trend of rolling bearing in practical engineering. With the rapid development of intelligent sensing technology and artificial intelligence, data-driven RUL prediction methods have been rapidly developed [9–11], main steps of these method contain: collecting bearing degradation data, building neural network model, performing feature extraction, and RUL prediction. For example, Wang et al [12] proposed a competitive temporal convolutional network (CTCN) to predict the RUL of rolling bearings, which used dual competitive attention and temporal convolutional networks to extract degradation features. Xu et al [13] proposed a multiscale temporal convolutional network (MSTCN) for RUL prediction, which combined multiscales with the self-attention mechanism to focus on the degradation features from both the temporal and spatial dimensions. Sun et al [14] combined the symmetric dot pattern (SDP) and composite multiscale permutation entropy (CMPE) to extract the bearing degradation features, which combined image and entropy to make the process of extracting degradation features simpler and more efficient. Qing et al [15] proposed a data-driven prediction method for bearings based on gated recurrent unit networks, a power spectrum related health indicator was constructed, and the gated recurrent unit network was utilised to predict the RUL of the bearings. In addition, some scholars have introduced physical knowledge into the models to construct more reliable prediction models. For example, Fu et al [16]

* Corresponding author at: College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China.
E-mail address: xqzhao@lut.edu.cn (X. Zhao).

https://doi.org/10.1016/j.measurement.2025.117278
Received 18 November 2024; Received in revised form 10 March 2025; Accepted 12 March 2025
Available online 15 March 2025

proposed a wavelet koopman predictor to achieve anomaly detection, which decomposed the non-stationary input into time-varying and time-invariant components by wavelet transform. Ma et al [17] proposed a time series prediction method based on Transformer and Kalman filtering, which embedded Kalman filtering into Transformer to achieve noise suppression. The above methods extract degradation features through neural networks. These features describe the degradation trend of rolling bearing more comprehensively and accurately. However, the above methods only carry out the study from the same working conditions. In other words, it is assumed that the training samples and test samples obey the same distribution. In practical engineering, most bearings operate under different operating conditions, and the difference in data distribution would lead to the sharp decrease for the RUL prediction performance [18–20].

RUL prediction under different operating conditions is the same concept as RUL prediction under cross domain. Cross domain refers to the process that a model transfers the learnt knowledge from the source domain to the target domain. To solve this problem, domain adaptive (DA) in transfer learning is introduced to the RUL prediction of rolling bearing. DA applies the learned source domain knowledge to the target domain by mining the similarity between the source and target domain features, which ultimately improves the RUL prediction accuracy of target domain [21,22]. For example, Li et al [23] proposed a knowledge-enhanced Transformer-based method for RUL prediction of rolling bearings, which used convolution and transformer to extract shared features in the source and target domains. Shuang et al [24] constructed a multi-branch network, multi-kernel maximum mean difference is used to reduce the distribution difference between the two domains. Zhao et al [25] constructed a feature extractor using residual network, and achieved RUL prediction. The above methods have achieved encouraging results in cross-domain RUL prediction for bearings. However, they still have the following shortcomings: (1) The variability of the distribution for rolling bearings at different stages is not fully considered. Rolling bearing can be classified into healthy state, slow degradation state and rapid degradation state, and the samples of different states have different discrepancies. It is difficult to mine the degradation features of rolling bearings at different stages using single-scale convolution. (2) The influence of critical features on the RUL prediction results for rolling bearings is not sufficiently considered. Rolling bearings often operate in noisy environments, and some critical degradation features can be drowned out by noise.

To address the above challenges, this paper proposes a wavelet decomposition and dynamic calibration domain adaptive network (DCDAN) to achieve RUL prediction for rolling bearings. DCDAN can capture the multi-scale features of rolling bearings at different stages of degradation and can effectively extract the features from different time scales by employing convolutional kernels of different scales, thus better adapting to the feature changes of rolling bearings in healthy, slow degradation and rapid degradation states. In addition, an attention mechanism is introduced to capture critical features under different stages, further improving the RUL prediction performance under cross domain. Specifically, to reduce the interference of noisy for vibration signals, the data enhancement module is designed, which consists of wavelet transform and frequency enhancement channel attention mechanism. Then, a multi-scale temporal convolution module (MTCM) and a broadcast multi-head self-attention gated recurrent unit (BMSA-GRU) are developed. MTCM not only achieves multi-scale feature extraction, but also enhances the model's ability to capture critical features by dynamically calibrating the weights. BMSA-GRU helps to capture long-term dependencies. Finally, MTCM and BMSA-GRU are utilized to construct the feature extractor, and the distributional differences between the source and target domain samples are mitigated by a multi-kernel maximum mean difference. The main contributions of this paper are described as follows:

(1) A data enhancement module is designed which consists of wavelet transform and frequency enhancement channel attention mechanism. The vibration signals are analyzed at multiple resolutions by discrete wavelet transform to obtain different levels of low and high frequency signals. The frequency channel attention mechanism is used to retain critical degradation features and reduce the missing frequency information.

(2) MTCM and BMSA-GRU are designed as the shared feature extractor. MTCM consists of several temporal channel attention blocks, each temporal channel attention block contains a dynamic temporal calibration unit and a multi-kernel efficient channel attention. The dynamic temporal calibration unit can adjust the weights of the convolution to fully exploit degraded features at different scales. BMSA-GRU further helps to capture the long-term dependencies of the time series.

(3) A smooth multi-kernel maximum mean difference is designed by introducing regularization, which further mitigates the difference in distribution between the samples in the source and target domains. Experimental results show that the proposed method is superior to existing RUL prediction methods and has good generalization performance.

The remainder of the paper is structured as follows: Section 2 presents related works. Section 3 describes the proposed methodology. Section 4 is case studies and analyses. Section 5 concludes.

2. Related works

2.1. Description of the problem

Transfer learning applies existing knowledge to an unknown domain. The domain to be transferred is called the source domain and the domain to be learnt is called the target domain in transfer learning. In this paper, we define the source domain as $\{x_i^s, y_i^s\}_{i=1}^{n_s}$, where n_s is the total number of samples in the source domain and y_i^s is the label corresponding to the samples. The sample of the source domain is χ^s that obeys a marginal probability distribution. We define the target domain as $\{x_i^t\}_{i=1}^{n_t}$, where n_t is the total number of samples in the target domain. The sample of the target domain is χ^t that obeys the marginal probability distribution. Since the working environments of rolling bearings change frequently, there is the big difference between the data distribution of the source domain and the target domain.

2.2. Separable convolution

Separable convolution consists of depth convolution and point-wise convolution. Specifically, first, deep convolution provides separate convolution operations for each channel of the input, its mathematical formula is described as follows:

$$Conv_{depth}(x, w^d) = \sum_{m=0}^{M} x_{(c,l+m)} \cdot w^d_{(c,m)} \tag{1}$$

where x denotes the input signal, w^d denotes the weight of the deep convolution, m denotes the convolution kernel, M denotes the total number of convolution kernels, c denotes the number of channels, and l denotes the length of the input signal.

Then, the output of the depth convolution is operated using point-wise convolution to fuse the information from the different channels, its mathematical formula is described as follows:

$$Conv_{point}(x^p, w^p) = \sum_{c=0}^{C} x^p_{(c,l)} \cdot w^p_{(c,C)} \tag{2}$$

where x^p denotes the input signal, w^p denotes the point-wise convolution weight, and C denotes the total number of channels.

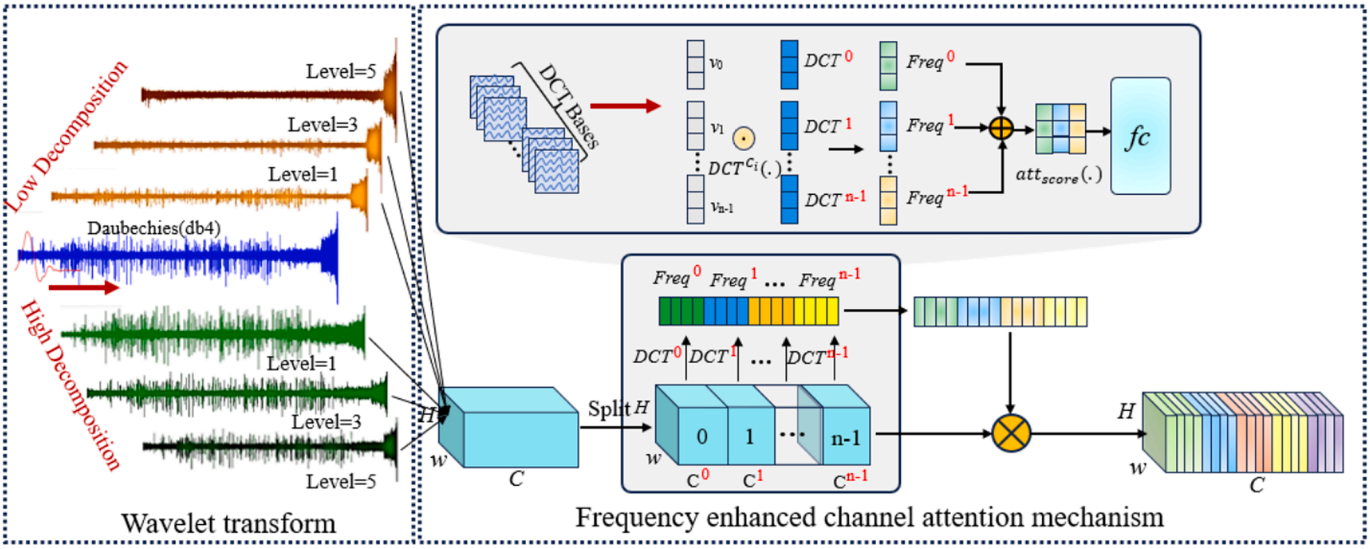


Fig. 1. Data enhancement module.

2.3. Dilated causal convolution

Dilated causal convolution consists of dilated convolution and causal convolution. Dilated convolution captures long-term dependencies by inserting the holes between the convolution kernels, allowing the convolution kernels to span larger time steps. Causal convolution ensures that the output depends only on the current and past inputs, avoiding leakage of future information. Thus, dilated causal convolution not only significantly improves the modelling capability, but also effectively reduces the computational burden.

2.4. Gated recurrent unit (GRU)

Gated Recurrent Unit (GRU) is a variant of recurrent neural network (RNN), which not only solves the problem of gradient vanishing during training of RNN, but also captures the long-term dependencies of time series. GRU consists of an update gate and a reset gate. The update gate controls the updating of the hidden state and helps the model to retain or forget the historical information. The reset gate controls how much the current moment is combined with the previous moment.

3. The proposed methodology

This section describes the main structure of the dynamic calibration domain adaptive network (DCDAN). In addition, the objective function is also presented for model training.

3.1. Dynamic calibration domain adaptive network

3.1.1. Data enhancement module

The data enhancement module consists of wavelet transform and frequency enhancement channel attention mechanism. Wavelet transform can effectively suppress noise interference in non-stationary signals [26]. The frequency enhancement channel attention mechanism not only enhances the critical features, but also obtains more frequency components by the discrete cosine transform (DCT) [27,28], which enhance useful features and avoid information missing by converting the signals from the time domain to the frequency domain. In addition, performing feature extraction in the frequency domain also helps to more effectively suppress noise in non-stationary signals. The structure of the data enhancement module is shown in Fig. 1.

In Fig. 1, the Daubechies (db4) wavelet is first chosen as the mother wavelet for 3-level wavelet transform. Specifically, we define the scale function and wavelet function in the wavelet transform, which are described as follows:

$$\kappa_{j,k}(t) = \sqrt{2}\sum_{n} h(n) \times \kappa(2t - n) \tag{3}$$

$$\phi_{j,k}(t) = \sum_{n} g(n) \times \phi(2t - n) \tag{4}$$

where $\kappa_{j,k}(t)$ is the scale function and $\phi_{j,k}(t)$ is the wavelet function, $h(n)$ is the low-pass filter and $g(n)$ is the high-pass filter. Therefore, the $J-$level DWT is described as follows:

$$x(t) = \sum_{k=0}^{2^{N-j}-1} \alpha_{j,k} 2^{-\frac{j}{2}} \times \kappa\left(2^{-j}t - k\right) + \sum_{j=1}^{J} \sum_{k=0}^{2^{N-j}-1} \beta_{j,k} 2^{-\frac{j}{2}} \times \phi\left(2^{-j}t - k\right) \tag{5}$$

where $x(t)$ denotes the vibration signal and $2n$ denotes the length of the vibration signal, $\alpha_{j,k}$ denotes the approximation coefficients sampled on the low-pass filter and $\beta_{j,k}$ denotes the approximation coefficients sampled off the high-pass filter. In order to avoid the duplication of degenerate information and reduce the computational burden, we select the wavelet reconstructed signals at Level = 1, Level = 3 and Level = 5 as the input signals for the model.

Then, the input is divided along the channel dimension, i.e., $[C^0, C^1, \ldots, C^{n-1}]$. For each channel, the frequency of DCT is described as follows:

$$Freq^i = DCT^{C^i}(x^i) = \sum_{l=0}^{L-1} x^i \cdot cos\left(\frac{\pi k}{L}\left(i + \frac{1}{2}\right)\right) \tag{6}$$

where C^i is the index of the channel, $C^i \in n \times L$. x^i is the input signal, which is obtained by discrete wavelet transform, $i \in \{0, 1, \ldots, L-1\}$. n is the number of channels and L is the length of the input signal. k is the frequency component index, $k \in \{0, 1, \ldots, L-1\}$.

Finally, each frequency component is fused and the attentional attention score is obtained by the sigmoid function, which is described as follows:

$$Freq = cat\left(Freq^0, Freq^1, \ldots, Freq^{n-1}\right) \tag{7}$$

$$att_{score} = \sigma(fc(Freq)) \tag{8}$$

where $cat(\cdot)$ is the fusion function, $\sigma(\cdot)$ is the sigmoid function and $fc(\cdot)$ is the fully connected layer. In the data enhancement module, different levels of low and high frequency signals are obtained by wavelet

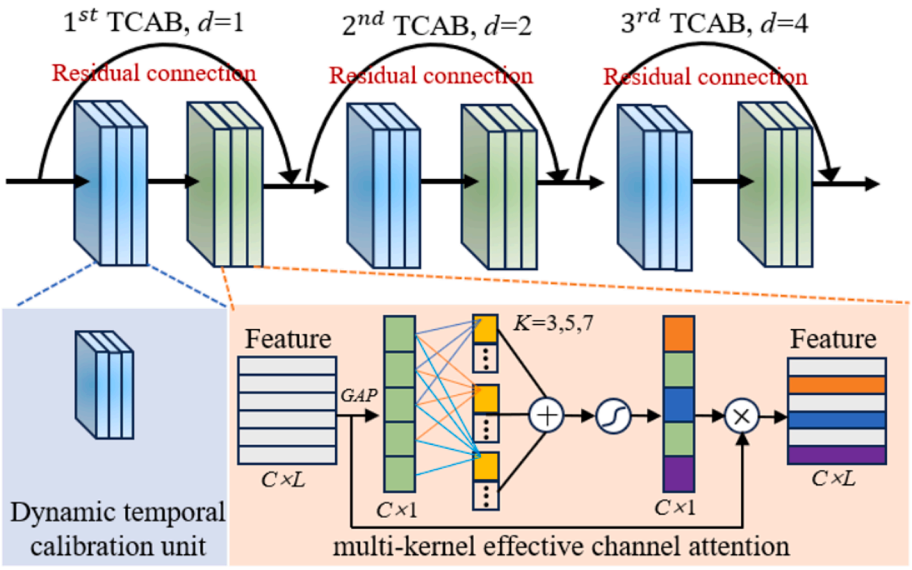


Fig. 2. Schematic structure for MTCM.

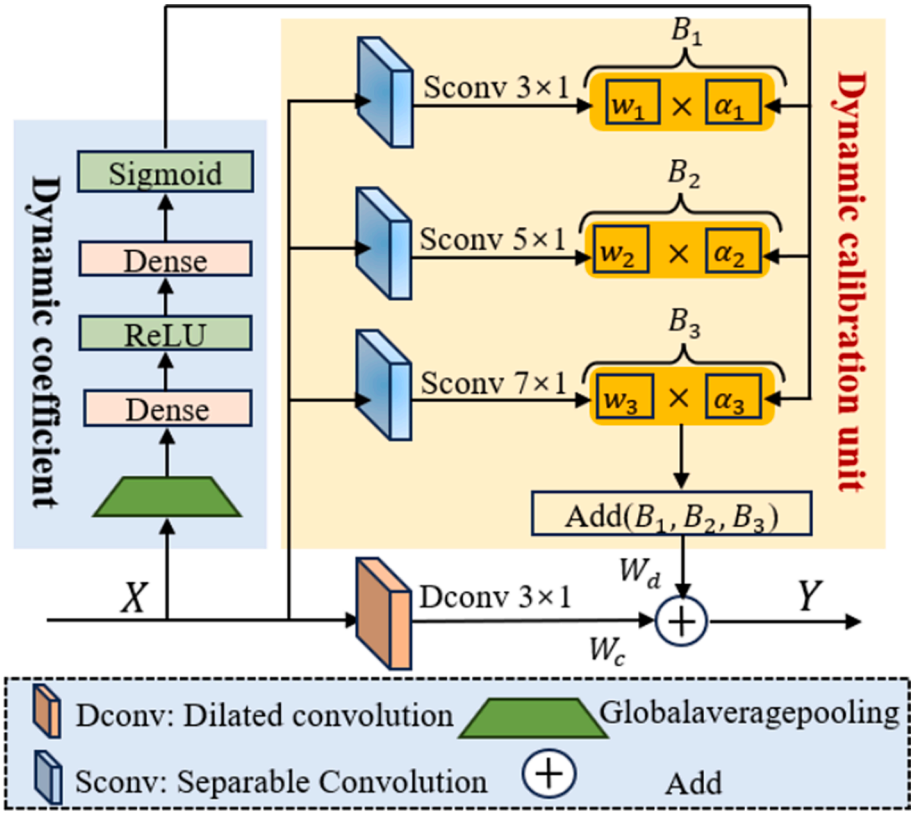


Fig. 3. Schematic diagram for DTCU.

transform and the useful degradation features are enhanced using frequency enhancement channel attention mechanism.

3.1.2. *Feature extractor*

When using domain adaptive for cross-domain RUL prediction of bearings, the metric function is usually used to reduce the difference between the source and target domains. However, there are large differences between the source and target domain samples. Only using the metric function to reduce the domain difference would result in poor cross-domain RUL prediction of bearings. Therefore, a dynamically calibrated feature extractor is designed to mine more domain-invariant features, which consists of a multi-scale temporal convolution module (MTCM) and a broadcast multi-head self-attention gated recurrent unit (BMSA-GRU). MTCM not only extracts rich information at different time scales, but also improves the receptive field without increasing parameters. BMSA-GRU further helps the feature extractor to capture the long-term dependencies of the time series. The details of MTCM and BMSA-GRU are described as follows.

(1) Multi-scale temporal convolution module

MTCM consists of several temporal channel attention blocks

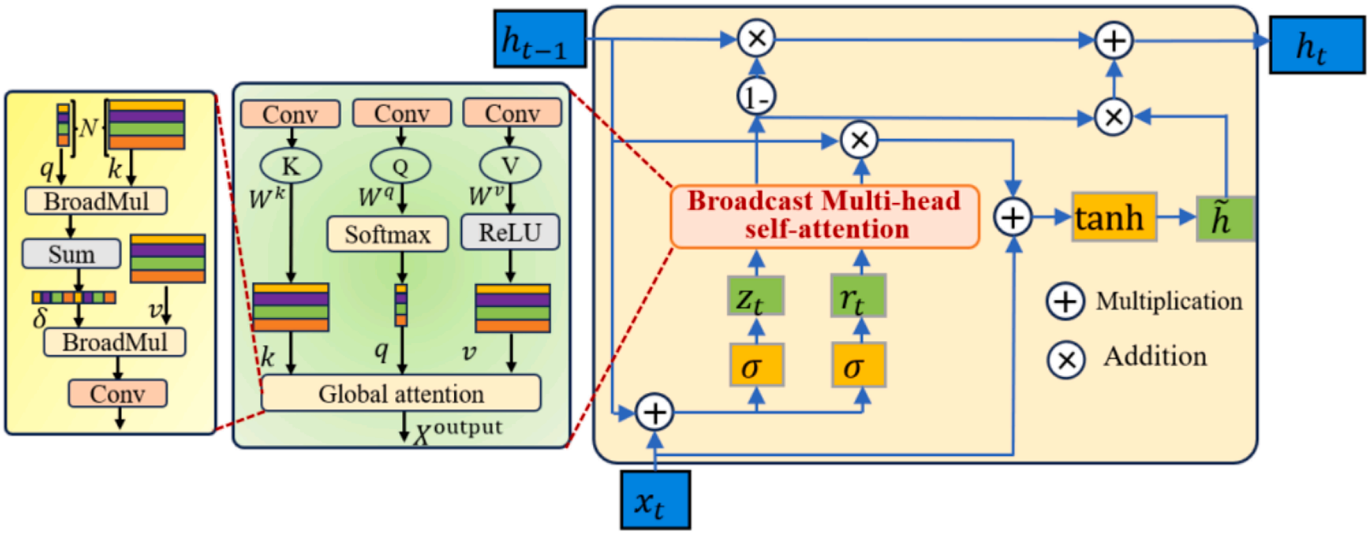


Fig. 4. Schematic diagram of BMSA-GRU.

(TCABs), each TCAB is connected to each other with residual connections for forward propagation of features. The dilation factor of the dilated convolution between different TCABs is set to $d = 1,2,3$, which enhances the receptive field of the feature extractor. The schematic diagram of MTCM is shown in Fig. 2. TCAB consists of dynamic temporal calibration unit (DTCU) and multi-kernel effective channel attention (MK-ECA). DTCU can assign different weights for different input samples to dynamically balance the weights, which enables the model to better learn degraded features. MK-ECA efficiently enhances the degraded features of the bearings and suppresses the redundant information through channel interacting encoding.

The schematic diagram of DTCU is shown in Fig. 3. DTCU is implemented by three steps.

Step1: The input of DTCU is X, $X \in L \times H$. The dynamic coefficient matrix is obtained by a global average pooling, two fully connected layers, a ReLU layer and a sigmoid layer. where ReLU layer is used to activate the features and ensure that the network captures important feature information. The sigmoid layer is used to generate the attention weights between 0 and 1, which is described as follows:

$$A = F_{sig}(F_F(F_R(F_F(F_{GAP}(X))))), \quad A \in [a_1, a_2, a_3] \tag{9}$$

where A is the matrix of dynamic coefficients, $F_{sig}(\cdot)$ is the sigmoid function, $F_F(\cdot)$ is the fully connected layer, $F_R(\cdot)$ is the ReLU function, and $F_{GAP}(\cdot)$ is the global average pooling layer.

Step2: Separable convolutions of different sizes are used for feature extraction. Specifically, the static weight coefficients of the input samples are obtained using separable convolutions of different sizes. The static weight coefficients are multiplied with the dynamic coefficient matrix to dynamically adjust the different weights of the input samples, which is described as follows:

$$W = F^d_{Sconv}(X), W \in [\omega_1, \omega_2, \omega_3] \tag{10}$$

$$B = A \times W, B \in [B_1, B_2, B_3] \tag{11}$$

where W is the static weight coefficient, B is the different weight matrices after adjustment, $F^d_{Sconv}(\cdot)$ is the separable convolution and d is the dilation factor.

Step3: The obtained different weight matrices are linearly summed. In addition, the dilated convolution is used to perform the convolution operation on the input X. The obtained results are fused with the weight matrices to achieve the output of DTCU, which is described as follows:

$$W_d = Add(B_1, B_2, B_3) \tag{12}$$

$$Y = W_d \oplus F^d_{Dconv}(X) \tag{13}$$

where $\oplus$ is the fusion operation, $F^d_{Dconv}(\cdot)$ is the dilated convolution, and d is the dilation factor.

(2) Broadcast multi-head self-attention gated recurrent unit

The multi-scale temporal convolution module can only extract the local degradation features. However, the feature extracting ability of the network is often limited under different operating conditions, resulting in poor prediction performance. GRU is not only capable of capturing long-term dependencies between different time steps, but also mitigates the problem of gradient vanishing [29]. Currently, GRU with attention mechanism only connects the output and input of the network together, which results in degradation features at different time steps not being captured. Thus, we use broadcast attention to construct the broadcast multi-head self-attention gated recurrent unit (BMSA-GRU) between the reset gate and the update gate, which the structure is shown in Fig. 4.

In Fig. 4, firstly, the reset and update gates of GRU control the missing of information and the degree of hidden layer information update, respectively. The update process for reset gate and update gate is described as follows:

$$z_t = \sigma\left(x_t w^z_x + h_{t-1} w^h_z + b_z\right) \tag{14}$$

$$r_t = \sigma\left(x_t w^r_x + h_{t-1} w^h_r + b_r\right) \tag{15}$$

where σ is the sigmoid function, w^z_x and w^r_x are the weight parameters corresponding to the input variables, w^h_z and w^h_r are the weight parameters corresponding to the hidden units. b_z and b_r are the bias parameters. In addition, the auxiliary hidden state of the input $\tilde{h}$ is described as follows:

$$h' = F_{\tanh}\left(x_t w^h_x + (r_t \otimes h_{t-1}) w^h_h + b_h\right) \tag{16}$$

where $F_{\tanh}$ is the tanh function, w^h_x and w^h_h are the weight parameters. b_h is the bias parameter, $\otimes$ is the elemental product operation.

Secondly, we use broadcast attention to construct the multi-head self-attention mechanism between the reset gate and the update gate. Broadcast attention not only can fully mine the global feature information, but also can effectively reduce the computational complexity by using convolution. Specifically, we define the query matrix Q, key matrix K and value matrix V, which are described as follows:

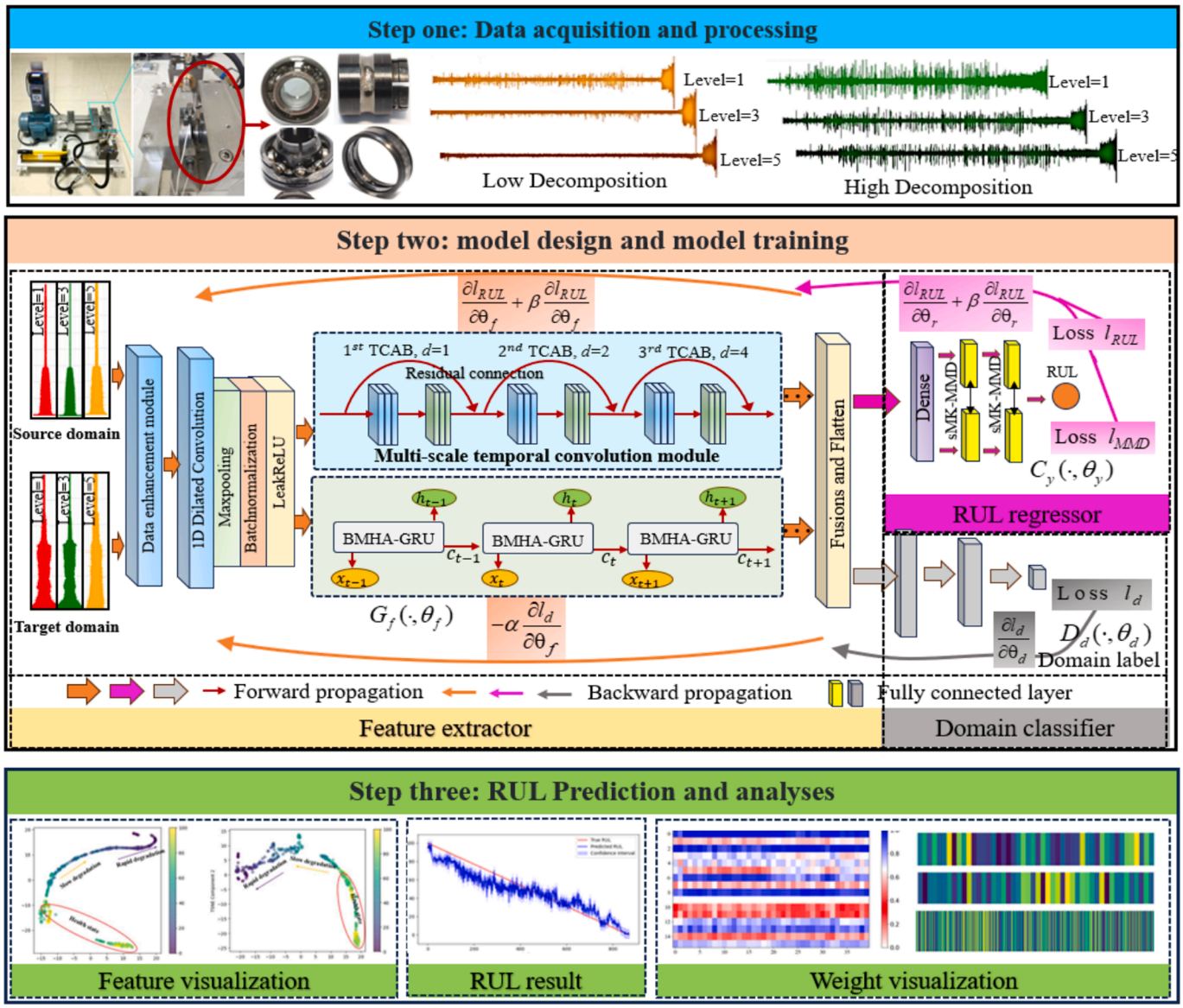


Fig. 5. Flowchart of RUL prediction for the DCDAN.

$$\begin{cases} Q = \mathrm{Softmax}(W^q \times \xi) \\ V = \mathrm{ReLU}(W^v \times \xi) \\ K = W^k \times \xi \end{cases} \tag{17}$$

where W^q, W^v and W^k are the convolution weights. W^q is used to calculate the contribution value at different time steps. The contribution value is assigned to W^k by broadcasting operation for adaptive selection of degraded features. The process is described as follows:

$$\delta = \sum_{i=1}^{N} \left(W_i^q \otimes W_i^k\right) \tag{18}$$

where $\otimes$ is the broadcast operation. δ is passed to the value matrix by the broadcast operation. Thus, the bearing degradation information is obtained from the global perspective. The process is described as follows:

$$X^{output} = \sum_{j=1}^{C} \left(\delta_i^b \otimes W_i^v\right) \tag{19}$$

where δ_i^b is the result of the broadcast operation performed on the query matrix and the key matrix, and C is the number of channels.

Finally, reset gate and update gate through multi-head self-attention mechanism are updated as follows:

$$z_t' = multihead\left(\sigma\left(x_t w_x^z + h_{t-1} w_z^h + b_z\right)\right) \tag{20}$$

$$r_t' = multihead\left(\sigma\left(x_t w_x^r + h_{t-1} w_r^h + b_r\right)\right) \tag{21}$$

where $multihead(\cdot)$ denotes the multi-head self-attention mechanism.

3.1.3. Smooth multi-kernel maximum mean difference

We design a smooth multi-kernel maximum mean difference by introducing regularization. Specifically, the Gaussian kernel is used as the basic kernel function, which is described as follows:

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{\theta^2}\right) \tag{22}$$

where θ is the bandwidth parameter of the kernel function. $\|\cdot\|^2$ is the square of the Euclidean distance. Then, the multicore maximum mean difference between the source and target domains is calculated, which is described as follows:

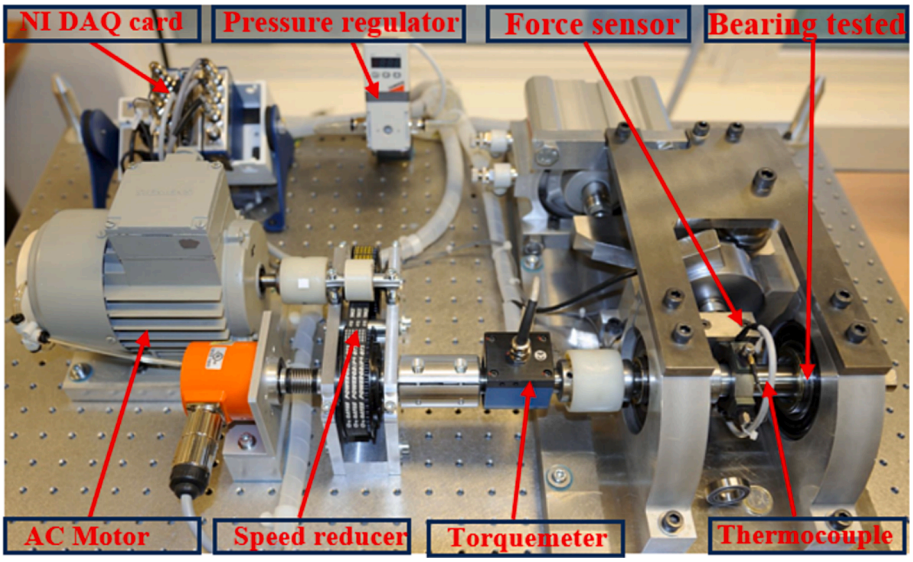


Fig. 6. FEMTO dataset of bearing degradation test rig.

Table 1
Division of the PHM2012 dataset.

Task	Condition	Training Source domain	Target domain	Test bearing
A	CS1 → CS2	IEEE1-1, IEEE1-2	IEEE2-2	IEEE2-1
B	CS1 → CS2	IEEE1-1, IEEE1-2	IEEE2-1	IEEE2-2
C	CS1 → CS2	IEEE1-1, IEEE1-2	IEEE2-1	IEEE2-4
D	CS2 → CS1	IEEE2-1, IEEE2-2	IEEE1-2	IEEE1-1
E	CS2 → CS1	IEEE2-1, IEEE2-2	IEEE1-1	IEEE1-2
F	CS2 → CS1	IEEE2-1, IEEE2-2	IEEE1-1	IEEE1-4

$$MKMMD^2(X,Y)=\sum_{j=1}^{p}w_j\left\|\frac{1}{n}\sum_{i=1}^{n}\phi(x_i)-\frac{1}{m}\sum_{j=1}^{m}\phi(x_j)\right\|_H^2 \tag{23}$$

where X and Y denote the source and target domains, respectively. w_j is the weight of the kernel function, $\phi(\cdot)$ is the feature mapping operation, n is the total number of samples in the source domain, and m is the total number of samples in the target domain. $\|\cdot\|_H^2$ is the L2 paradigm of the difference between the source and target domains. Finally, the smooth of the multi-kernel maximum mean difference is achieved by introducing the regular term in Eq. (23), which is described as follows:

$$sMKMMD^2(X,Y)=\sum_{j=1}^{p}w_j\left\|\frac{1}{n}\sum_{i=1}^{n}\phi(x_i)-\frac{1}{m}\sum_{j=1}^{m}\phi\left(y_j\right)\right\|_H^2+\gamma\sum_{j=1}^{p}w_j^2 \tag{24}$$

where γ is the regularization factor, which is obtained by cross-validation.

Table 2
Division of the XJTU dataset.

Task	Condition	Training Source domain	Target domain	Test bearing
1	CS1 → CS2	XJTU1-1, XJTU 1–2	XJTU 2–2	XJTU 2–5
2	CS1 → CS2	XJTU 1–1, XJTU 1–2	XJTU 2–5	XJTU 2–2
3	CS2 → CS1	XJTU 2–1, XJTU 2–2	XJTU 1–2	XJTU 1–5
4	CS2 → CS1	XJTU 2–1, XJTU 2–2	XJTU 1–5	XJTU 1–2

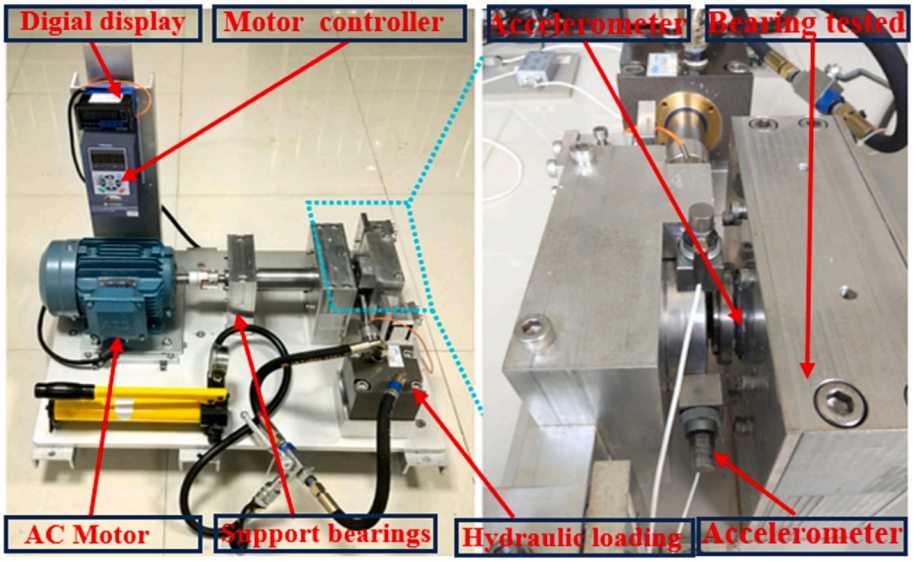


Fig. 7. XJTU-SY dataset of bearing degradation test rig.

Table 3
Parameters of DCDAN.

Names	Layers	Kernel/Hidden size	Activation	Input	Output
Feature extractor	Input	/	/	/	2560 × 6
	Data enhancement module	/	/	2560 × 6	2560 × 6
	Convolution	12	LeakyReLU (0.1)	2560 × 6	2560 × 16
	Max-pooling	2	/	1280 × 16	640 × 16
	1st TCAB	3	/	640 × 16	320 × 12
	2nd TCAB	3	/	320 × 12	160 × 8
	3rd TCAB	3	/	160 × 8	160 × 6
	BMSA-GRU	6	/	640 × 16	640 × 12
	Fusion	/	/	/	18 × 1
Domain discriminator	Convolution	3	ReLU	2560 × 6	1280 × 32
	Max-pooling	2	/	1280 × 32	640 × 32
	Convolution	3	ReLU	640 × 32	320 × 64
	Max-pooling	2	/	320 × 64	160 × 64
	Dense	128	ReLU	160 × 64	128 × 1
	Dense	1	Sigmoid	128 × 1	1
Regressor	Dense	64	ReLU	18 × 1	64 × 1
	Dense	1	Linear	64 × 1	1

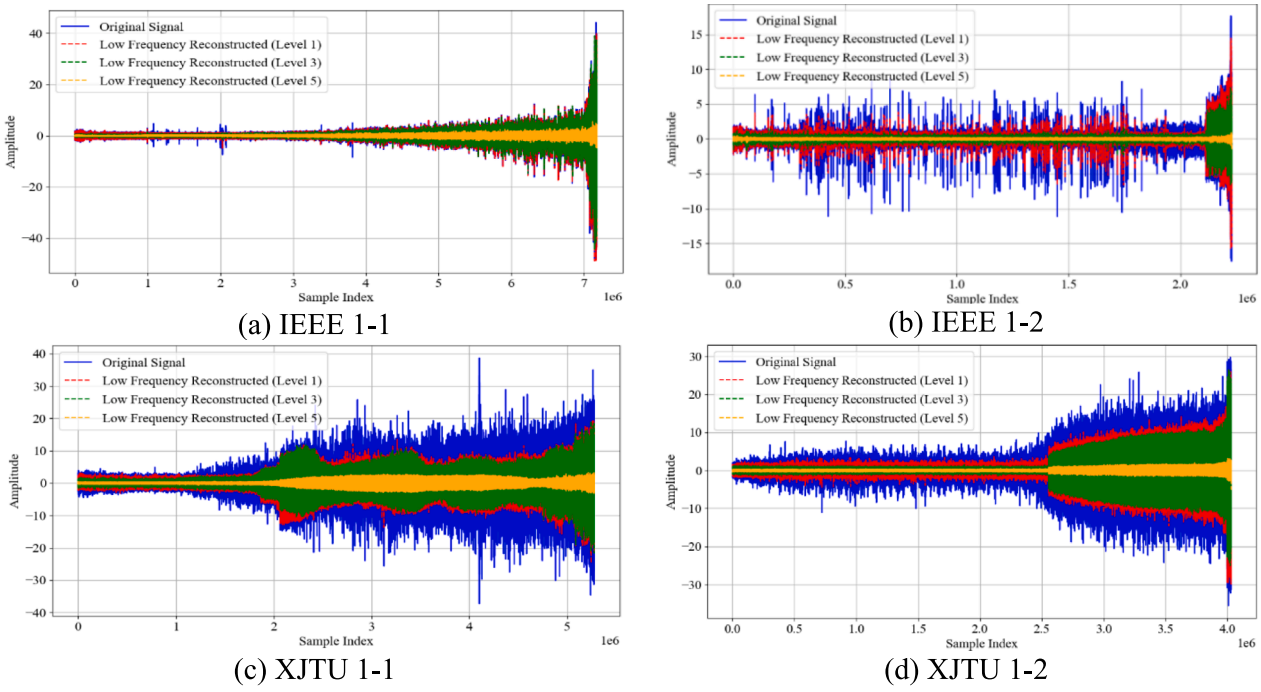


Fig. 8. Reconstruction of low frequency signal after wavelet decomposition.

3.2. Model optimization function

The optimization function of DCDAN consists of the following objective functions: the maximizing domain adversarial loss L_{DA}, the minimizing RUL prediction loss L_{RUL} and the minimizing sMK-MMD loss $L_{sMK-MMD}$, which are described as follows:

$$L_{DA} = -\left(\frac{1}{n_s}\sum_{i=1}^{n_s}\log\left[D_d\left(f_s^i,\theta\right)\right]+\frac{1}{n_t}\sum_{i=1}^{n_t}\log\left[D_d\left(f_t^i,\theta\right)\right]\right) \tag{25}$$

where n_s is the source domain samples, n_t is the target domain samples, $D_d\left(f_s^i,\theta\right)$ is the predicted label for the source domain, and $D_d\left(f_t^i,\theta\right)$ is the predicted label for the target domain.

$$L_{RUL} = \frac{1}{n_s}\sum_{i=1}^{n_s}\left(y_i^{strue}-y_i^s\right)^2 \tag{26}$$

where y_i^{strue} is the true RUL and y_i^s is the predicted RUL.

$$L_{sMK-MMD} = \left\|E[\phi(x_s)]-E[\phi(x_t)]\right\|_H^2 \tag{27}$$

where $E(\cdot)$ is the mathematical expectation, $\phi(\cdot)$ is the regenerated Hilbert space mapping and H is the feature kernel. Thus, the overall optimized loss function of DCDAN is described as follows:

$$L_{total} = L_{RUL} - \mu L_{DA} + \lambda L_{sMK-MMD} \tag{28}$$

where μ and λ are the weight parameters for domain adversarial loss and multi-kernel maximum mean loss, respectively. The parameters of DCDAN are optimized by using Adam, which are described as follows:

$$\begin{aligned}\theta_f &\leftarrow \theta_f - \eta\left(\frac{\partial L_{RUL}}{\partial\theta_f}-\mu\frac{\partial L_{DA}}{\partial\theta_f}+\lambda\frac{\partial L_{sMK-MMD}}{\partial\theta_f}\right)\\ \theta_r &\leftarrow \theta_r - \eta\frac{\partial L_{RUL}}{\partial\theta_r}\\ \theta_d &\leftarrow \theta_d - \eta\frac{\partial L_{DA}}{\partial\theta_d}\end{aligned} \tag{29}$$

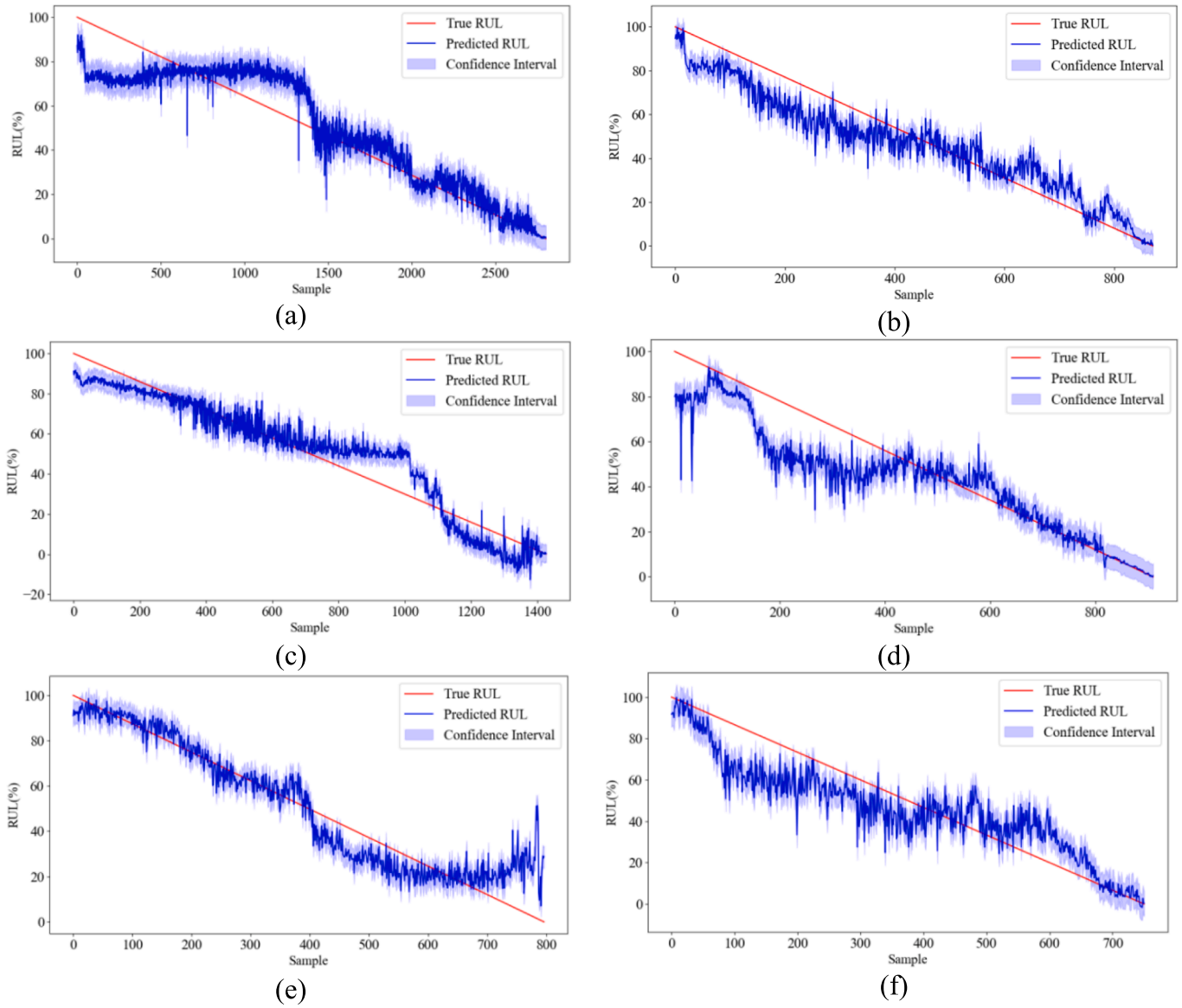


Fig. 9. RUL prediction results for the six tasks of the PHM 2012 dataset. (a) TaskA (b) TaskB (c) TaskC (d) TaskD (e) TaskE (f) TaskF.

where η is the learning rate. Algorithm 1 shows the training and testing process for DCDAN.

Algorithm 1: Training and Test Procedures for DCDAN

Inputs: source domain $\{x_i^s, y_i^s\}_{i=1}^{n_s}$, target domain $\{x_i^t\}_{i=1}^{n_t}$
1: Set hyperparameters such as convolution layer, activation, learning rate, batch size, epoch
2: Initialise the weights and biases for DCDAN
3: Divide the training and test of the source and target domains
4: **IF** DCDAN does not reach convergence **do**
5: **For** each training epoch **do**
6: Calculate the output of RUL regressor
7: Solve L_{RUL} based on Eq. (26)
8: Calculate the outputs for domain classifiers and sMK-MMD
9: Solve L_{DA} based on Eq. (25)
10: Solve $L_{sMK-MMD}$ based on Eq. (27)
11: Update model parameters according to Eq. (29)
12: **End for**
13: Save the trained model
Outputs: Test data are loaded into the trained DCDAN to obtain prediction results of the test samples

3.3. The RUL prediction framework for DCDAN

The overall flowchart of DCDAN is shown in Fig. 5, which can be summarized in three steps:

Step 1: Data acquisition and processing. The vibration signals are collected under variable working conditions. Then, the vibration signals are decomposed into multiple levels by discrete wavelet transform to obtain the reconstructed signals at level = 1, level = 3 and level = 5. The RUL labels are created for the reconstructed source domain samples.

Step 2: Model design and training. A dynamic calibration domain adaptive network is built and initialized, which consists of a feature extractor, a domain classifier and a RUL regressor. The multi-scale temporal convolution module and broadcast multi-head self-attention gated recurrent unit are embedded into the feature extractor to mine more domain-invariant features. In addition, the distribution difference between the source and target domains is further reduced by multi-kernel maximum mean difference in the RUL regressor. The network is trained using some labelled samples at the source domain and unlabelled samples at the target domain, and the network parameters are updated by an overall optimized loss function.

Step 3: The test samples at the target domain are fed into the trained

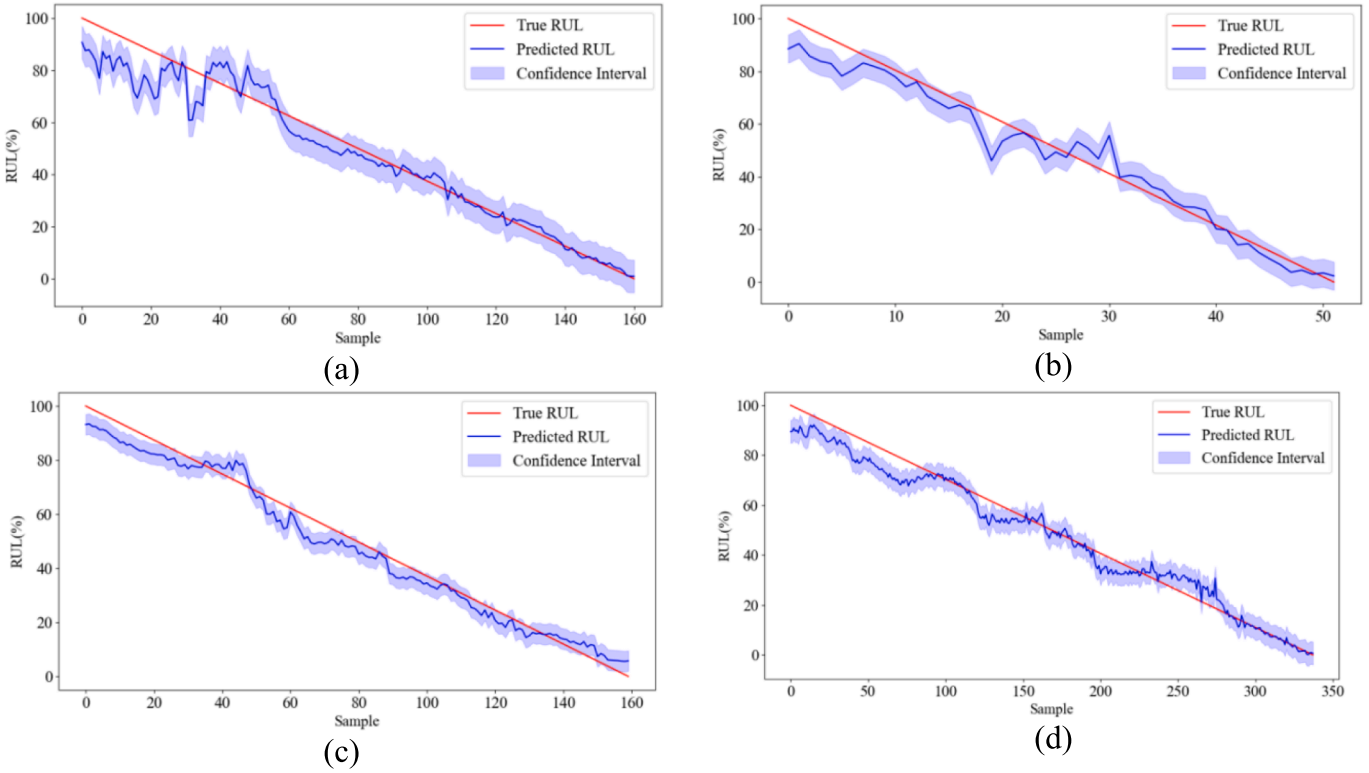


Fig. 10. RUL prediction results for the six tasks of the XJTU dataset. (a) Task1 (b) Task2 (c) Task3 (d) Task4.

network for RUL prediction.

4. Experiments and analysis

In order to validate the prediction performance of DCDAN, two datasets are used for case studies in this paper. The computer configuration for the experiments is AMD Riptide 5-4600H processor and the experimental framework is keras = 2.0.0.

4.1. Introduction to the datasets

4.1.1. IEEE PHM 2012 dataset

The IEEE PHM 2012 dataset is derived from the PRONOSTIA test rig [30]. The platform consists of an AC motor, the tested rolling bearing, and an acceleration sensor, as shown in Fig. 6. The dataset contains full life cycle acceleration data of 17 bearings under three different operating conditions. The acquisition of the vibration signals is stopped when the amplitude of the vibration signals exceeds 20 g. The sampling frequency of the vibration signals is 25.6 KHz, and the data are recorded once every 10 s. 2560 data samples are collected each time. In the experiments, we select the full life acceleration degradation under working conditions (1) and (2). The details are shown in Table 1.

4.1.2. XJTU-SY bearing dataset

The XJTU-SY bearing dataset is obtained from Xi'an Jiaotong University, and the experimental platform is shown in Fig. 7 [31]. The tested bearing of this experimental platform is LDK UER204, and the acceleration data of 15 bearings under three working conditions are collected by hydraulic loading system. The sampling frequency is 25.6 KHz and 32,768 data samples are collected each time. In the experiments, we select the bearing data under working conditions (1) and (2) for cross domain experiments. The details are shown in Table 2.

4.1.3. Evaluation metrics

We use the mean absolute error (MAE), root mean square error (RMSE) [32,33], and scoring function to evaluate the RUL prediction performance [34]. The MAE and RMSE formulas are calculated as follows:

$$Er_i = \left(Pre_i^{rul} - Act_i^{rul}\right)/Pre_i^{rul} \tag{30}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|Pre_i^{rul} - Act_i^{rul}\right| \tag{31}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Pre_i^{rul} - Act_i^{rul}\right)^2} \tag{32}$$

where N is the total number of samples, Pre_i^{rul} is the predicted RUL at time i, and Act_i^{rul} is the true RUL at time i.

The mathematical formulation of the scoring function is described as follows:

$$Score = \alpha\frac{1}{k}\sum_{i=1}^{k}A_i + \beta\frac{1}{n-k}\sum_{k+1}^{n}A_i \tag{33}$$

where A_i denotes the weighted error between the true RUL and the predicted RUL at time i, α and β are the weighting factors.

$$A_i = \begin{cases} e^{-\ln(0.6)\bullet(Er_i/10)}, Er_i \le 0 \\ e^{\ln(0.6)\bullet(Er_i/40)}, Er_i > 0 \end{cases} \tag{34}$$

where e is the exponential function.

4.2. Model parameters and data processing

4.2.1. Model parameters

The parameters of DCDAN are shown in Table 3. The feature extractor consists of MTCM and BMSA-GRU. The domain discriminator consists of three fully connected layers. The regressor consists of two fully connected layers. The layer choice of wavelet transform and the

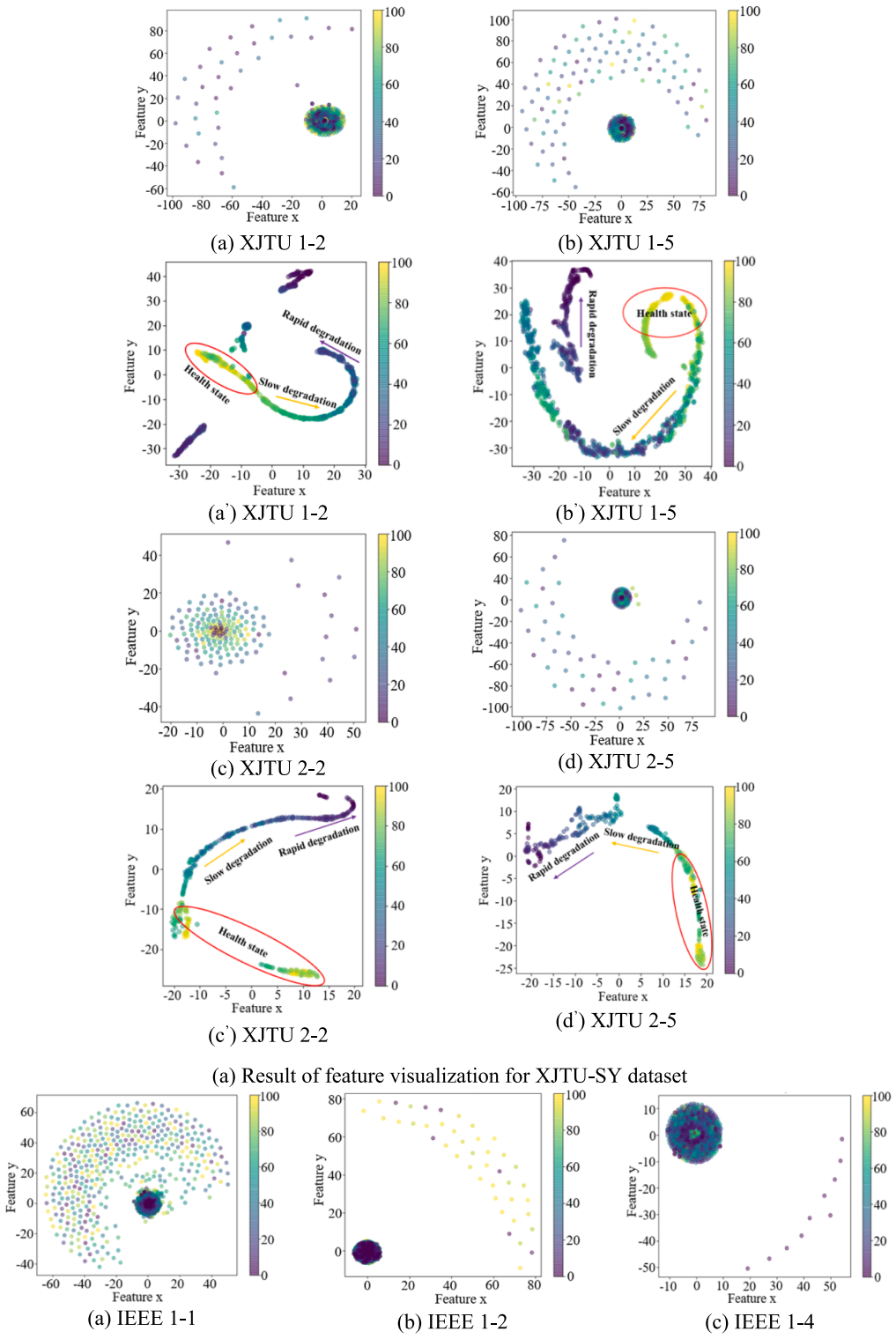
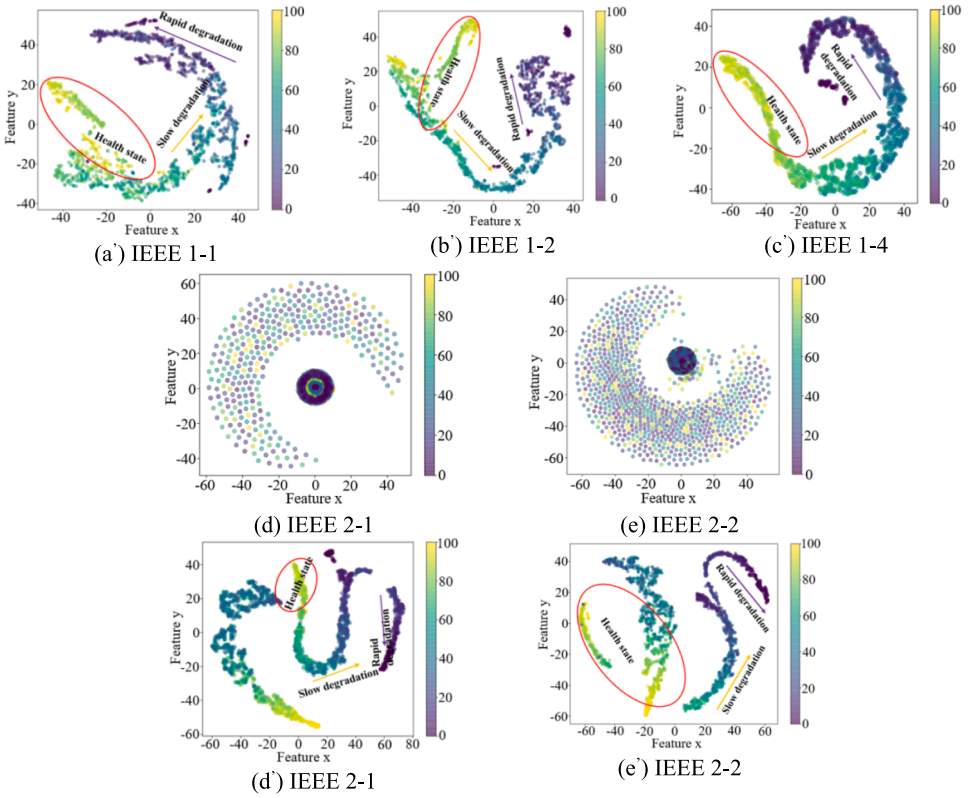


Fig. 11. Result of feature visualisation after t-SNE.

hyperparameters of the model are obtained through cross-validation experiments. The network parameters are updated by Adam optimizer. The learning rate is 0.001 and the batch size is 128. The training network is 200 epoch.

4.2.2. Data processing

The neural network has the strong feature extraction ability, which can directly extract the degradation features from the vibration signals of the rolling bearings. However, due to the harsh working environment, the vibration signals are interfered to become non-stationary by noise. The wavelet transform can decompose the original vibration signals into low-frequency signals and high-frequency signals, thus effectively filtering out the noise interference. Taking bearing 1-1, 1-2 in IEEE 2012 and XJTU as an example, we plot the low-frequency of 3-level



(b) Result of feature visualization for IEEE PHM 2012 dataset

Fig. 11. (*continued*).

Table 4
Prediction results of different methods for PHM 2012 dataset.

Methods	Metrics	TaskA	TaskB	TaskC	TaskD	TaskE	TaskF	Average
DCDAN	MAE	8.9269	6.7631	10.610	8.2963	8.0121	7.2457	8.3090
	RMSE	12.570	9.4395	12.997	10.740	9.7236	8.7619	10.705
DTM	MAE	13.602	16.087	13.277	18.394	14.046	10.711	14.352
	RMSE	17.017	19.987	17.261	20.116	18.052	13.633	17.677
MAOR	MAE	9.6880	11.327	13.100	12.855	9.1294	11.904	11.334
	RMSE	12.752	13.685	17.448	14.710	11.263	13.337	13.865
TCN-RSCB	MAE	15.271	17.487	14.546	16.174	13.948	13.576	15.167
	RMSE	18.943	21.701	18.109	17.221	15.690	16.116	17.963
BTACN	MAE	19.698	17.354	13.425	19.641	16.191	19.458	17.627
	RMSE	23.310	19.876	17.378	23.691	20.106	21.467	20.971
MCNN	MAE	14.955	17.078	13.542	16.702	14.256	17.443	15.662
	RMSE	20.624	22.918	18.511	18.433	17.464	19.908	19.643
SCTA	MAE	15.381	22.822	11.944	19.327	16.724	15.399	16.932
	RMSE	18.142	25.081	15.558	21.722	18.525	16.944	19.328

decomposition in Fig. 8. In Fig. 8, wavelet transform can effectively filter the noise interference for the non-stationary signals, and the filtering effect is more obvious with the increase of decomposition layers.

In order to convert vibration signals of different amplitudes under the same standard, we perform normalization on the vibration signals after wavelet decomposition. The process is described as follows:

$$X = \frac{x - \varphi}{\xi}, \quad \varphi = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \xi = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \varphi)^2} \tag{35}$$

where φ is the mean value of the input data and ξ is the standard deviation of the input data. The scale differences between the vibration signals can be eliminated by normalization, which helps to be more stable and improves the training speed of the model.

4.3. RUL prediction results and visualization analysis

4.3.1. RUL prediction results for rolling bearings

Fig. 9 is the RUL prediction results for the six tasks of the PHM 2012 dataset, which shows the true RUL values in red and the predicted RUL values in blue. All tasks are repeated ten times to reduce randomness. We conclude the following by observing the plots: (1) The predicted RUL values of DCDAN have a good downward trend in the plots, and the predicted values basically fluctuate at the true RUL. (2) The predicted values for the remaining tasks are closer to the true values in the later stages of the prediction, except for the anomaly for TaskE. This is due to the fact that the data for TaskE are collected during the rapid degradation failure of the bearing. Therefore, it exhibits stronger nonlinearity in the later stages of equipment operation, which leads to a larger difference between the distributions in the source and target domain, and

Table 5
Prediction results of different methods for XJTU dataset.

Methods	Metrics	Task1	Task2	Task3	Task4	Average
DCDAN	MAE	4.6426	4.3146	3.8637	3.9257	4.1866
	RMSE	6.6624	5.7107	4.4152	5.0238	5.4530
DTM	MAE	7.7290	8.0233	6.4101	6.3877	7.1375
	RMSE	9.9411	10.722	7.6077	8.8656	9.2841
MAOR	MAE	6.8980	5.4331	6.3870	7.0798	6.4494
	RMSE	8.0055	7.4960	8.0454	8.1074	7.9135
TCN-RSCB	MAE	13.707	14.950	9.4857	10.307	12.112
	RMSE	15.674	16.594	11.417	12.342	14.006
BTACN	MAE	12.571	13.091	10.025	11.665	11.838
	RMSE	13.910	14.872	12.377	13.291	13.612
MCNN	MAE	8.4244	9.4592	7.4266	11.502	9.2030
	RMSE	10.059	11.374	9.5114	13.223	11.041
SCTA	MAE	11.470	10.865	13.257	9.3107	11.225
	RMSE	15.951	12.560	14.955	11.334	13.700

ultimately affects the prediction accuracy of TaskE. (3) The predicted values of DCDAN have large fluctuations in the initial stage, which are caused by abnormal vibrations of the bearings during the run-in stage. In the later stages of prediction, DCDAN can extract and capture more domain invariant features, which makes the predicted values closer to the true values.

Fig. 10 shows the RUL prediction results for the four tasks of the XJTU dataset, the experimental setup is the same as the PHM 2012 dataset, we have the following findings: (1) DCDAN obtains good prediction results in all four tasks, the predicted RUL has a similar decreasing trend with the true RUL and follows the true RUL values well. (2) For Task1, the predicted RUL fluctuates around the true RUL in the early stage. However, the predicted values can follow the true values well in the later stages, because the samples of the target domain and the source domain differ greatly, and the multi-kernel maximum mean difference cannot effectively reduce the distributional difference between the two domains. However, with the continuous training of the model, the dynamic temporal calibration unit can dynamically balance the weights of the model, which effectively reduces the domain differences and improves the prediction accuracy. (3) In the later stages of prediction, the prediction accuracy of DCDAN is better than that of the earlier stages.

4.3.2. Data feature visualization

We observe that DCDAN has good predictive performance from the previous section. However, due to the black-box character of deep learning, it is difficult to understand the inner workings of the model. To further visualise the feature extraction capability, we use the t-SNE technique to visualise the raw data and extracted high-dimensional features, as shown in Fig. 11. Each point represents the extracted features of DCDAN, and the colour of each point indicates the label of RUL. The degradation features extracted are mixed at the initial layer by DCDAN and the degradation trend of the bearings cannot be revealed. However, with the increase of the feature extraction capability, the extracted deep features show three different stages of the state: the healthy state, the slow degradation state, and the rapid degradation state. This indicates that DCDAN can mine more domain invariant features and has good RUL prediction performance.

4.4. Comparison with other methods

We select six state-of-the-art RUL prediction methods to analyse the cross-domain RUL prediction tasks. The evaluation metrics are described in Section 4.1.3. The comparison methods include domain-based adversarial methods [18,35], TCN-based models [36,37] and CNN/RNN-based models [38,39]. In addition, in order to fully assess the overall performance of the proposed method, we calculate the average values of the metrics. The results are shown in Table 4 and Table 5, it can be observed that DCDAN has the lowest average MAE value and average RMSE value. Specifically, compared to the second ranked MAOR, the

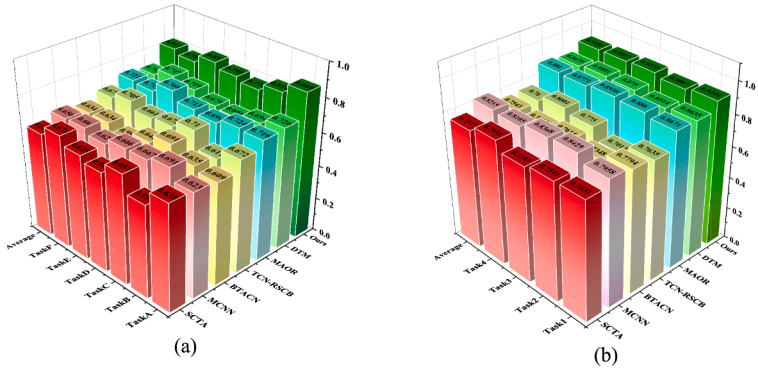


Fig. 12. Score values of different methods. (a) PHM 2012 dataset (b) XJTU dataset.

Table 6
Predicted results of ablation experiments.

Methods	Metrics	TaskA	TaskB	TaskC	Task1	Task2	Task3
DCDAN	MAE	8.9269	6.7631	10.610	4.6426	4.3146	3.8637
	RMSE	12.570	9.4395	12.997	6.6624	5.7107	4.4152
WDEM	MAE	17.495	12.497	18.714	8.1330	8.5896	10.642
	RMSE	19.621	15.603	21.087	10.477	10.727	12.430
WDTCU	MAE	15.966	10.933	15.518	7.2975	6.9232	7.8468
	RMSE	17.613	12.667	17.277	9.7795	9.0137	9.2964
WMKCA	MAE	12.519	9.2973	13.902	8.2738	5.6875	6.9580
	RMSE	14.170	11.021	15.412	10.917	8.4765	8.0102
WBMSA	MAE	14.004	10.697	14.725	7.0001	16.191	7.6215
	RMSE	16.865	11.140	16.208	9.4165	7.9293	9.2941
WMKMMD	MAE	10.732	9.4460	13.195	6.3763	6.0497	5.0210
	RMSE	12.639	11.017	14.954	8.0593	7.5958	6.4888

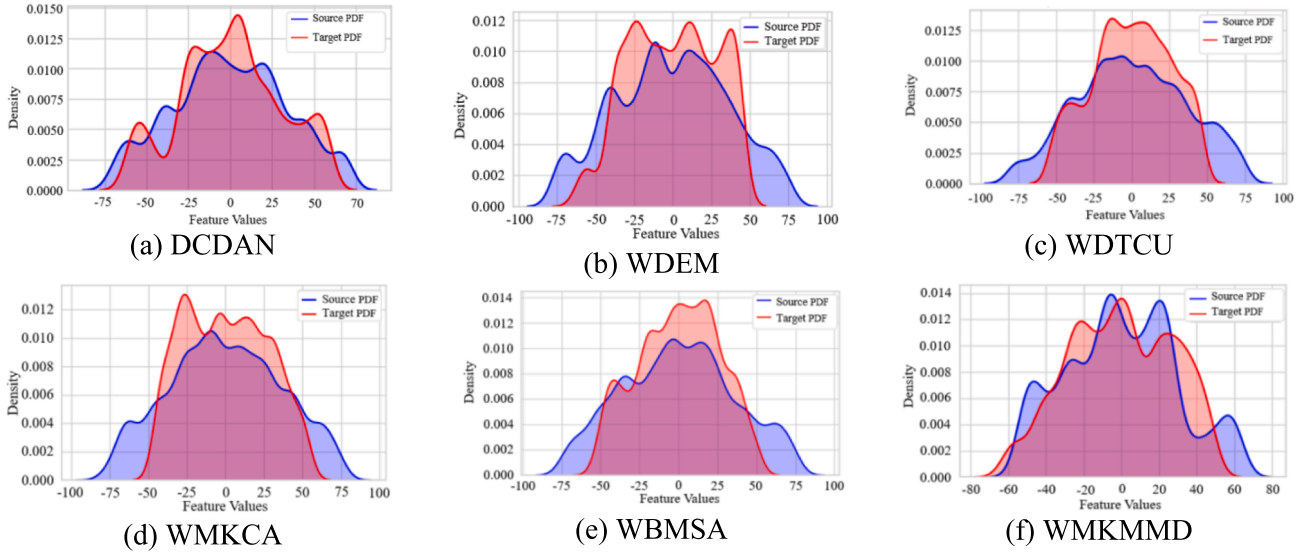


Fig. 13. Adaptation effects of ablation experiments.

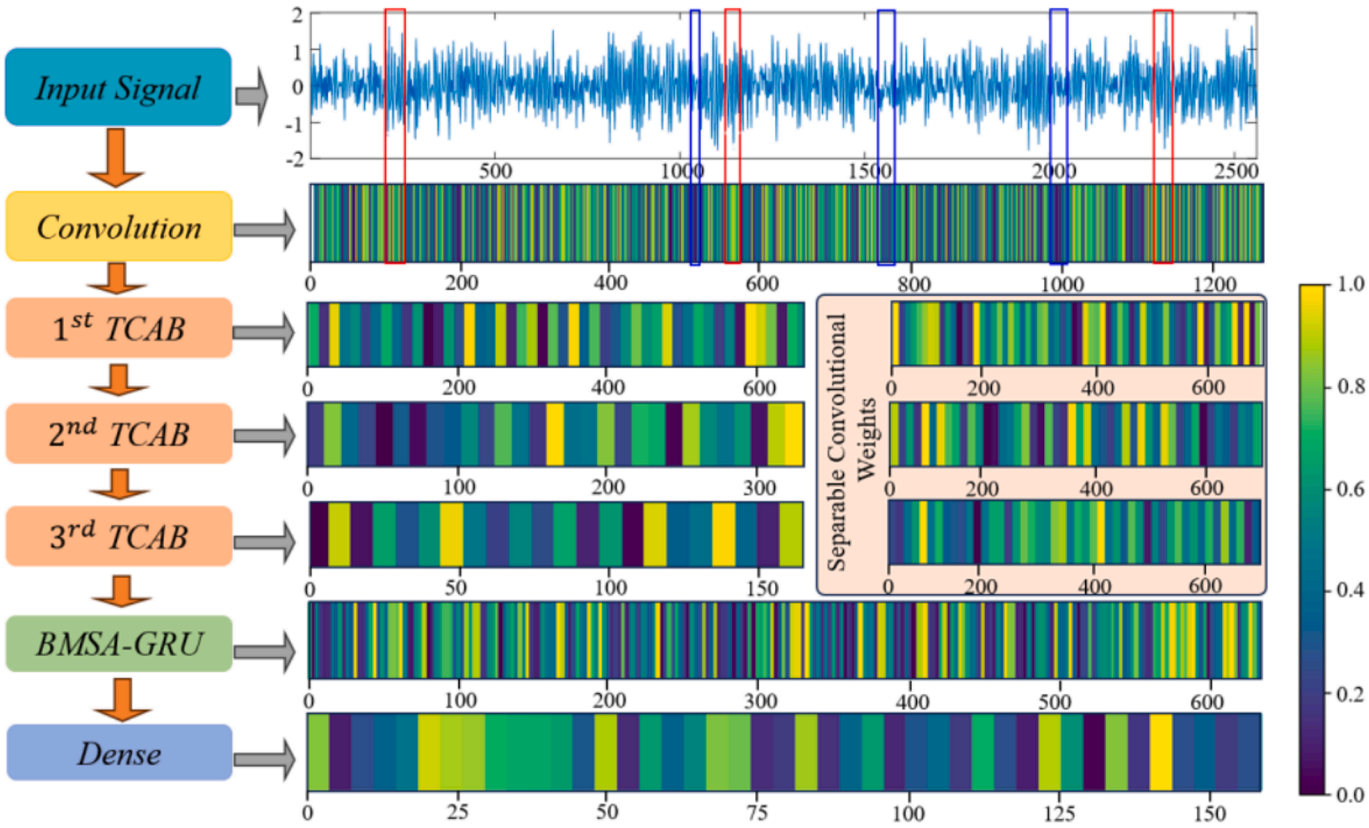


Fig. 14. Visualization results of channel weight averages for DCDAN.

proposed method improves the average MAE and average RMSE by 3.025 and 3.16. Compared to the second ranked MAOR, DCDAN improves the average MAE and average RMSE by 2.2628 and 2.4605. The above experimental results show that DCDAN has superior performance in cross-domain RUL prediction. The reason is that the noise interference in bearing vibration signals can be eliminated by the data enhancement module. In addition, the designed MTCM and MHSA-GRU have strong feature learning ability, which can effectively mine the domain-invariant features between the source and target domains. Finally, the distribution difference between the source and target domains is reduced by using the multi-kernel maximum mean difference, which results in the satisfactory RUL prediction accuracy.

In order to further visualise the score values of the compared methods, we plot the 3D bar chart as shown in Fig. 12. From Fig. 12, we have the following findings: (1) In the cross-domain RUL prediction task across different datasets, the bar heights are the highest and the corresponding score values are also the largest for DCDAN. Based on the definition of the Score function, this indicates that the RUL prediction values of DCDAN can follow the rolling bearings degradation trend well in the later stages of RUL prediction. (2) The overall Score values of DTM

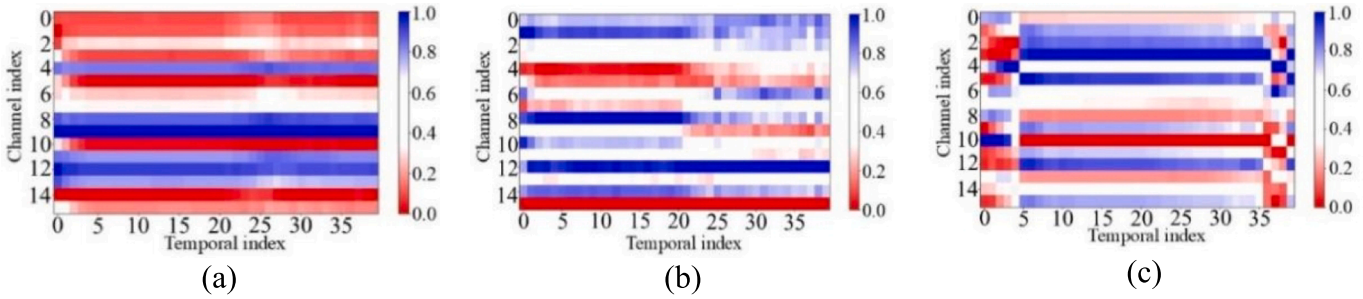


Fig. 15. Results of feature visualization after attentional mechanisms.

and MAOR are better than that of other comparison experiment for methods. This is because DTM and MAOR use the domain adaptive, which can reduce the distribution difference between the source and target domains and improve the RUL prediction accuracy.

4.5. Ablation experiments and interpretability analysis

4.5.1. Ablation experiments

In order to comprehensively analyse the effectiveness of each component for DCDAN, this section performs ablation experiments. Specifically, they contain the following: without data enhancement module (WDEM), without dynamic temporal calibration unit (WDTCU), without multi-kernel effective channel attention (WMKCA), without broadcast multi-head self-attention gated recurrent unit (WBMSA), and without multi-kernel maximum mean difference (WMKMMD).

The parameter settings of the ablation experiments are the same as the proposed method and each experiment is repeated ten times, as shown in Table 6. From Table 6, it can be observed as follows: (1) Data enhancement module has the greatest impact on the accuracy of cross-domain RUL prediction. Without data enhancement module, the average MAE and RMSE values of the six tasks are reduced by 6.1576 and 6.3576 compared with the proposed method, respectively. This further indicates that the noise interference can be effectively reduced by the data enhancement module, which enables DCDAN to better learn the bearing degradation features and improve the RUL prediction accuracy. (2) The effects of DTCU and BMSA-GRU for DCDAN are flagrant. Without DTCU and BMSA-GRU, the proposed method cannot fully mine the domain-invariant features, resulting in lower RUL prediction accuracy for DCDAN. (3) In cross-domain RUL prediction, the smooth multi-kernel maximum mean difference can reduce the difference between the source and target domain during the model training process, thus improving the RUL prediction accuracy.

4.5.2. Feature adaptive visualization

To further visualise the effect of each module in DCDAN, we use TaskA for probability density figure (PDF) visualization as shown in Fig. 13. In Fig. 13(a), both the source and target domains of DCDAN obey the same distribution, and their PDF curves show the largest overlapping region. However, in Fig. 13(b), the PDF curves of the source and target domains have different distributions and the overlap region is also minimum. This indicates that the data enhancement module not only reduces noise interference but also effectively captures critical degradation features. In Fig. 13(c) and Fig. (e), the PDF curves of the source and target domains are similar, which indicates that the temporal calibration unit and the multi-head self-attention mechanism have the same effect in the model, which pays more attention to the domain-invariant features.

4.5.3. Model weights visuzlisation

To visually explore the DCDAN feature learning process, we select the first sample of IEEE 1–1 data for visual analysis. Fig. 14 shows the average values of the channel weights for each layer. The weights of each layer are normalized and different colours represent different weight values. Brighter colour means higher weight value. In Fig. 14, the initialization layer (convolutional layer) can assign higher weights for the regions with higher amplitudes and lower weights for the regions with lower amplitudes. In addition, the trend of the highlights in the weight visualization plot is almost consistent with the impulse signals of the input samples. For example, the red boxes are the higher amplitude regions and the blue boxes are the lower amplitude regions. This indicates that the initialization layer of DCDAN can effectively focus on the degraded features and make the good foundation for extracting more transferable features. Then, the critical degradation features are highlighted after three TCAB layers. For example, there are fewer and fewer regions that are given higher weights in the 3rd TCAB. This shows that the TCAM layer can further mine the degraded critical features and filter out some redundant information. In addition, BMSA-GRU layer has the same function as the initialization layer. However, BMSA-GRU can further focus on critical degradation features due to the introduction of the broadcast multi-head self-attention mechanism. Finally, the critical degradation features are further highlighted through the dense layer. In conclusion, DCDAN can extract more transferable features and focus on critical degraded features by weight visualization.

4.5.4. Attention mechanism visualization

To further illustrate the effects of multi-kernel efficient attention and broadcast attention, we plot the feature maps of multi-kernel efficient attention and broadcast attention, as shown in Fig. 15. Fig. 15(a) shows the feature map of the attention mechanism, and Fig. 15(b) shows the feature map of multi-kernel efficient attention. In Fig. 15(b), multi-kernel efficient attention can give different weights to the feature information of different channels from the channel dimension, thus highlighting the bearing degradation information. Fig. 15(c) is the feature map after the action of broadcast attention, we can see the weight changes are mainly in the beginning and ending regions. This indicates that broadcast attention can weigh the feature information from the time dimension, thus focusing on the effect of different time steps on the RUL prediction.

5. Conclusion

In this paper, a dynamic calibration domain adaptive network (DCDAN) for rolling bearing RUL prediction is proposed. First, a data enhancement module is designed, which consists of wavelet transform and frequency enhancement channel attention mechanism. The original vibration signals are decomposed by wavelet transform to obtain different levels of low and high frequency signals. In addition, the critical features are enhanced by frequency enhancement channel attention mechanism. Then, a dynamic calibration domain adaptive network is designed, which the feature extractor consists of MTCM and BMSA-GRU. MTCM dynamically adjusts the weights to help the network extract fine-grained features at different degradation stages. BMSA-GRU can establish long-term dependency and improve the RUL prediction accuracy. Finally, the low-frequency and high-frequency signals are fed into

DCDAN to achieve RUL prediction for rolling bearings. The experimental results show that DCDAN has stronger generalization and better prediction accuracy.

CRediT authorship contribution statement

Yazhou Zhang: Writing – review & editing, Writing – original draft, Methodology. **Xiaoqiang Zhao:** Writing – review & editing, Methodology, Funding acquisition. **Zhenrui Peng:** Writing – review & editing, Funding acquisition. **Rongrong Xu:** Validation, Methodology. **Yongyong Hui:** Writing – review & editing, Methodology.

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.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (No.62263021), the College Industrial Support Project of Gansu Province (2023CYZC-24), the Science and Technology Project of Gansu Province (24JRRA172), and the Outstanding Postgraduate Innovation Star Project of Gansu Provincial Department of Education (2025CXZX-491)

Data availability

Data will be made available on request.

References

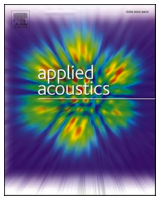
[1] C. Wu, J. He, W. Shen, W. Xu, S. Liu, Remaining useful life prediction across operating conditions based on deep subdomain adaptation network considering the weighted multi-source domain, Knowl.-Based Syst. 301 (2024) 112291.
[2] Y. Shang, X. Tang, G. Zhao, P. Jiang, T.R. Lin, A remaining life prediction of rolling element bearings based on a bidirectional gate recurrent unit and convolution neural network, Measurement 202 (2022) 111893.
[3] Q. Yang, B. Tang, Q. Li, X. Liu, L. Bao, Dual-frequency enhanced attention network for aircraft engine remaining useful life prediction, ISA Trans. 141 (2023) 167–183.
[4] W. Mao, J. Liu, J. Chen, X. Liang, An interpretable deep transfer learning-based remaining useful life prediction approach for bearings with selective degradation knowledge fusion, IEEE Trans. Instrum. Meas. 71 (2022) 1–16.
[5] L. Lin, J. Wu, S. Fu, S. Zhang, C. Tong, L. Zu, Channel attention & temporal attention based temporal convolutional network: A dual attention framework for remaining useful life prediction of the aircraft engines, Adv. Eng. Inf. 60 (2024) 102372.
[6] M. Furqon, M. Pratama, L. Liu, H. Habibullah, K. Dogancay, Mixup domain adaptations for dynamic remaining useful life predictions, Knowl.-Based Syst. 295 (2024) 111783.
[7] J. Shi, J. Zhong, Y. Zhang, B. Xiao, L. Xiao, Y. Zheng, A dual attention LSTM lightweight model based on exponential smoothing for remaining useful life prediction, Reliab. Eng. Syst. Saf. 243 (2024) 109821.
[8] Y. Zhang, X. Zhao, R. Xu, Z. Peng, A dual-stream temporal convolutional network for remaining useful life prediction of rolling bearings, Meas. Sci. Technol. 36 (1) (2024) 016206.
[9] W. Lu, Y. Wang, M. Zhang, J. Gu, Physics guided neural network: Remaining useful life prediction of rolling bearings using long short-term memory network through dynamic weighting of degradation process, Eng. Appl. Artif. Intel. 127 (2024) 107350.
[10] J. Sun, X. Zhang, J. Wang, Lightweight bidirectional long short-term memory based on automated model pruning with application to bearing remaining useful life prediction, Eng. Appl. Artif. Intel. 118 (2023) 105662.
[11] Y. Duan, X. Cao, J. Zhao, M. Li, X. Yang, A spatio-temporal fusion autoencoder-based health indicator automatic construction method for rotating machinery considering vibration signal expression, IEEE Sens. J. (2023).
[12] W. Wang, et al., A novel competitive temporal convolutional network for remaining useful life prediction of rolling bearings, IEEE Trans. Instrum. Meas. (2023).
[13] Z. Xu, Y. Zhang, Q. Miao, An attention-based multi-scale temporal convolutional network for remaining useful life prediction, Reliab. Eng. Syst. Saf. (2024) 110288.
[14] Y. Sun, Z. Wang, Remaining useful life prediction of rolling bearing via composite multiscale permutation entropy and Elman neural network, Eng. Appl. Artif. Intel. 135 (2024) 108852.
[15] Q. Ni, J. Ji, K. Feng, Data-driven prognostic scheme for bearings based on a novel health indicator and gated recurrent unit network, IEEE Trans. Ind. Inf. 19 (2) (2022) 1301–1311.
[16] L. Fu, M. Ma, Z. Zhai, Deep koopman predictors for anomaly detection of complex IOT systems with time series data, IEEE Internet Things J. (2024).
[17] M. Ma, L. Fu, Z. Zhai, R.-B. Sun, Transformer based Kalman Filter with EM algorithm for time series prediction and anomaly detection of complex systems, Measurement 229 (2024) 114378.
[18] K. Liu, Y. Li, Remaining useful life prediction across machines using multi-source adversarial online knowledge distillation, Eng. Appl. Artif. Intel. 130 (2024) 107726.
[19] M. Zeng, F. Wu, Y. Cheng, Remaining useful life prediction via spatio-temporal channels and transformer, IEEE Sens. J. (2023).
[20] Y. Ding, P. Ding, M. Jia, A novel remaining useful life prediction method of rolling bearings based on deep transfer auto-encoder, IEEE Trans. Instrum. Meas. 70 (2021) 1–12.
[21] Y. Ding, P. Ding, X. Zhao, Y. Cao, M. Jia, Transfer learning for remaining useful life prediction across operating conditions based on multisource domain adaptation, IEEE/ASME Trans. Mechatron. 27 (5) (2022) 4143–4152.
[22] J. Yan, Z.-S. Ye, S. He, Z. He, A feature disentanglement and unsupervised domain adaptation of remaining useful life prediction for sensor-equipped machines, Reliab. Eng. Syst. Saf. 242 (2024) 109736.
[23] Y. Li, J. Li, H. Wang, C. Liu, J. Tan, Knowledge enhanced ensemble method for remaining useful life prediction under variable working conditions, Reliab. Eng. Syst. Saf. 242 (2024) 109748.
[24] L. Shuang, X. Shen, J. Zhou, H. Miao, Y. Qiao, G. Lei, Bearings remaining useful life prediction across equipment-operating conditions based on multisource-multitarget domain adaptation, Measurement (2024) 115026.
[25] C. Zhao, X. Huang, S. Li, Y. Li, L. Sun, A new domain adaption residual separable convolutional neural network model for cross-domain remaining useful life prediction, ISA Trans. 145 (2024) 239–252.
[26] J. Kim, S. Sin, J. Kim, Early remaining-useful-life prediction applying discrete wavelet transform combined with improved semi-empirical model for high-fidelity in battery energy storage system, Energy 297 (2024) 131285.
[27] L. Xiang, H. Bing, X. Li, A. Hu, A frequency channel-attention based vision Transformer method for bearing fault identification across different working conditions, Expert Syst. Appl. 262 (2025) 125686.
[28] M. Jiang, P. Zeng, K. Wang, H. Liu, W. Chen, H. Liu, FECAM: Frequency enhanced channel attention mechanism for time series forecasting, Adv. Eng. Inf. 58 (2023) 102158.
[29] Y. Qin, D. Chen, S. Xiang, C. Zhu, Gated dual attention unit neural networks for remaining useful life prediction of rolling bearings, IEEE Trans. Ind. Inf. 17 (9) (2020) 6438–6447.
[30] W. Mao, J. Chen, J. Liu, X. Liang, Self-supervised deep domain-adversarial regression adaptation for online remaining useful life prediction of rolling bearing under unknown working condition, IEEE Trans. Ind. Inf. 19 (2) (2022) 1227–1237.
[31] L. Jiang, T. Zhang, W. Lei, K. Zhuang, Y. Li, A new convolutional dual-channel Transformer network with time window concatenation for remaining useful life prediction of rolling bearings, Adv. Eng. Inf. 56 (2023) 101966.
[32] M. Miao, J. Yu, Z. Zhao, A sparse domain adaption network for remaining useful life prediction of rolling bearings under different working conditions, Reliab. Eng. Syst. Saf. 219 (2022) 108259.
[33] Y. Cao, Y. Ding, M. Jia, R. Tian, A novel temporal convolutional network with residual self-attention mechanism for remaining useful life prediction of rolling bearings, Reliab. Eng. Syst. Saf. 215 (2021) 107813.
[34] Y. Wang, L. Deng, L. Zheng, R.X. Gao, Temporal convolutional network with soft thresholding and attention mechanism for machinery prognostics, J. Manuf. Syst. 60 (2021) 512–526.
[35] J. Zhuang, Y. Cao, M. Jia, X. Zhao, Q. Peng, Remaining useful life prediction of bearings using multi-source adversarial online regression under online unknown conditions, Expert Syst. Appl. 227 (2023) 120276.
[36] Y. Zhang, X. Zhao, Remaining useful life prediction of bearings based on temporal convolutional networks with residual separable blocks, J. Braz. Soc. Mech. Sci. Eng. 44 (11) (2022) 527.
[37] H. Liang, J. Cao, X. Zhao, Multi-sensor data fusion and bidirectional-temporal attention convolutional network for remaining useful life prediction of rolling bearing, Meas. Sci. Technol. 34 (10) (2023) 105126.
[38] J. He, C. Wu, W. Luo, C. Qian, S. Liu, Remaining useful life prediction and uncertainty quantification for bearings based on cascaded multi-scale convolutional neural network, IEEE Trans. Instrum. Meas. (2023).
[39] H. Tian, L. Yang, B. Ju, Spatial correlation and temporal attention-based LSTM for remaining useful life prediction of turbofan engine, Measurement 214 (2023) 112816.

Applied Acoustics 240 (2025) 110934

Contents lists available at ScienceDirect

Applied Acoustics

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



An interpretable frequency-enhanced domain adaptive network for cross-domain fault diagnosis of rotating machinery

Yazhou Zhang [a], Xiaoqiang Zhao [a,b,*], Zhenrui Peng [a,b], Yongyong Hui [a,b], Rongrong Xu [a], Peng Chen [c]

[a] College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China
[b] Gansu Key Laboratory of Advanced Control of Industrial Processes, Lanzhou 730050, China
[c] College of Electrical and Electronic Engineering, Lanzhou Petrochemical University of Technology, Lanzhou 730050, China

ARTICLE INFO

Keywords:
Fault diagnosis
Rotating machinery
Domain adversarial network
Hierarchical frequency attention mechanism
Frequency-enhanced residual block

ABSTRACT

The performance of domain adaptive fault diagnosis methods can be limited under noisy environments and variable operating conditions, and existing domain adaptive methods lack interpretability. Therefore, to address the above issues, an interpretable frequency-enhanced domain adaptive network (IFEDAN) for cross-domain fault diagnosis of rotating machinery is proposed in this paper. First, the time-domain signals are converted into the frequency domain using the Fast Fourier Transform (FFT) to enhance the representation of frequency-domain fault features. Additionally, the Morlet wavelet is introduced in the initial layer of the model for weight initialization, which enhances the model's ability to capture fault features. Then, a frequency-enhanced residual block is designed, which not only helps the model to capture more transferable features, but also further enhances the useful features from both local and global perspectives. Finally, Entropy Maximum Mean Difference (EMMD) loss is designed. EMMD uses the entropy value to determine the bandwidth of the Gaussian kernel, which enhances the stability of the training and decision boundaries. Validation is performed on the public dataset, the roller gear (RG) dataset and the Lanzhou University of Technology (LUT) dataset. The results show that IFEDAN has excellent cross-domain diagnostic performance. When performing cross-domain diagnosis between different datasets, the average diagnosis accuracy of IFEDAN reaches 93.81 %, which is higher than the comparison methods.

1. Introduction

In modern industry, rotating machinery is developing towards high precision and intelligence. However, the failure or even accident of rotating machinery has become more frequent due to high-intensity work, which puts forward higher requirements for reliability and safety [1–3]. Gearboxes and bearings, as key components of rotating machinery, their health status will affect the normal operation of the whole rotating machinery. Therefore, it is of great significance to carry out monitoring and fault diagnosis for rotating machinery to ensure its safe operation [4–6].

In the past, researchers have done many encouraging studies for fault diagnosis, including signal processing, mechanism analysis, and shallow network methods [7,8]. However, the vibration signals of rotating machinery are nonlinear and nonstationary, and methods that rely on expert knowledge and shallow networks present great challenges. In recent years, with the development of sensor technology and artificial intelligence, intelligent diagnostic methods based on data-driven have become a hot spot for research [9–12]. Particularly, diagnostic methods based on the framework of CNN have developed rapidly and catered to the problems of diagnosis with small samples, strong noise environments and class imbalance [13,14]. For example, Zhang et al. [15] proposed a fault diagnosis method based on wavelet denoising and KANTransformer, which achieves fault diagnosis under strong noise environments and small sample conditions. Lv et al. [12] proposed a rolling bearing fault diagnosis method based on adaptive feature decomposition and Transformer. Two conditions must be met for the above methods to achieve superior diagnostic performance: (1) The training samples and the test samples should be under the same load conditions. That is to say that the training and test samples follow the same data distribution. (2) A large number of labeled training samples are required to complete the model training. However, in real industrial

* Corresponding author at: College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China.
E-mail address: xqzhao@lut.edu.cn (X. Zhao).

https://doi.org/10.1016/j.apacoust.2025.110934
Received 10 March 2025; Received in revised form 25 June 2025; Accepted 5 July 2025

production, rotating machinery usually work under different loads. Different distributions exist for the training samples and test samples, which leads to the diagnostic results of the model deviating from expectations.

To address the above problems, cross-domain diagnosis methods based on domain adaptation (DA) have attracted widespread attention from researchers [16,17]. Cross-domain diagnosis refers to fault diagnosis tasks conducted between the data with different feature distributions [18,19]. Cross-domain diagnosis involves source domain and target domain. The source domain has the data with sufficient labelled information, which is often used to train model. The data in the target domain are used to test model and has less labelled information. DA can be classified into distance metric-based and adversarial training-based methods for fault diagnosis [20,21]. Distance metric-based methods map learned features to a shared space, and use statistical methods to calculate the distribution distance. For example, Wang et al. [22] proposed a federated transfer rolling bearing fault diagnosis method, which uses a dynamic filter to filter out poor source domain labels and constructs better pseudo-labels for the target domain. Pu et al. [23] proposed a single-domain incremental generative network for fault identification under variable speed. The method generates multiple augmented domains from a single source domain dataset, and further improves the augmented domains using a generator. Han et al. [24] proposed a two-classifier adaptive transfer diagnostic model, which reduces the negative transfer between the two domains. Furthermore, some scholars have also proposed semi-supervised and unsupervised domain adaptation methods and achieved encouraging results. For example, Yan et al. [25] proposed an unsupervised fault detection method based on Transformer and graph convolutional network, which utilized Transformer for global modelling and graph convolutional network to extract domain-invariant features. Jiang et al. [26] proposed a semi-supervised fault diagnosis method based on multi-sensor data fusion, which assigned different weights to sensor data through adaptive learning and performed feature extraction to achieve mechanical fault diagnosis. Song et al. [27] proposed a contrast-assisted domain-specificity removal network for semi-supervised fault diagnosis. Yan et al. [28] proposed a fault detection method based on multi-modal imitation learning, which achieved unsupervised learning of multi-modal data using a transferable imitation learning network.

The above DA performs domain-invariant feature extraction from the perspective of parameter transfer. Currently, DA domain adaptation methods based on adversarial training are also receiving widespread attention from scholars. DA based on adversarial training utilises the idea of domain adversarial networks to reduce domain differences. [29]. For example, Yu et al. [30] proposed a deep convolutional and self-attention network to achieve fault identification for rolling bearing, and the domain differences are reduced by multi-kernel maximum mean difference. Jiang et al. [31] proposed a correlation-aligned multi-adversarial adaptive network based on correlation alignment for fault classification of rotating machinery, which uses separable convolution and residual concatenation to improve the feature extractor. Coral distance is used to align the subdomain features and global features, which ultimately achieves the extraction and identification of transferable features. In addition, since the fault information of gearboxes and bearings occurs in a periodic form, researchers have tried to incorporate the attention mechanism into the feature extractor to capture more transferable features. For example, Yao et al. [32] proposed a hierarchical adversarial multi-objective domain approach for gearbox fault diagnosis. The method develops a three-level domain adaptive strategy. Each level enhances the domain invariant features using a different attention mechanism to improve the cross-domain diagnostic performance. Shao et al. [33] used self-attention to focus on globally relevant features, which improved the diagnostic performance under variable speed.

Although the above methods reduce the domain differences from adversarial training perspective and achieve the expected results in cross-domain diagnosis, the performance of the domain-adaptive diagnosis methods will still be limited under noisy environments or variable working conditions. In addition, the attention mechanisms improve the diagnostic performance by giving higher weights, but they are deficient in guidance from prior knowledge, resulting in the lack of interpretability for the above method. Therefore, some scholars have tried to improve the interpretability of methods by embedding physical information. For example, Cheng et al. [34] proposed a physics and data-driven Fourier neural operator network, which enhanced the network's interpretability by directly embedding physical information into the neural network. Han et al. [35] proposed an interpretable deep feature fusion network, which guided the attention mechanism to focus on important features using wavelet knowledge. Cheng et al. [36] proposed a rolling bearing fault diagnosis method combining fault feature enhancement with blind deconvolution, which integrated empirical wavelet transform to identify fault features for rolling bearings.

Inspired by the above interpretability methods, this paper proposes an interpretable frequency-enhanced domain adaptive network (IFEDAN) for cross-domain fault diagnosis of rotating machinery. Unlike the above interpretability methods, IFEDAN introduces Morlet wavelet into the initial layer of the model for weight initialization, which uses prior knowledge-based initial states to improve the interpretability. IFEDAN enhances the domain representation in three dimensions: input samples, feature extraction, and distance metric. First, the time-domain signal is converted to the frequency domain by Fast Fourier Transform (FFT) to improve the frequency-domain fault feature representation, and Morlet wavelet are introduced at the initial layer of the model for the initial of the weights. Then, a frequency-enhanced residual block is designed. This block can fully extract the low-frequency and high-frequency fault features to more comprehensively handle the input signals and improve the accuracy of fault identification. Finally, the improved MMD (EMMD) enhances the decision boundary and improves the generalization performance. EMMD performs the bandwidth selection of the Gaussian kernel through the entropy value, which avoids the problem of instability during training. The main contributions of this paper are as follows:

(1) An interpretable frequency-enhanced domain adaptive network is proposed, which uses Morlet wavelet for weight initialization in the first layer of the feature extractor. It enhances the learning of transferable features by introducing physical knowledge and improves the interpretability and robustness.
(2) A frequency-enhanced residual block is designed, which consists of group normalization (GN), PReLU, a hierarchical frequency attention mechanism and residual connections. GN and PReLU reduce the effect of noise and enhance the feature learning ability for model. In addition, a hierarchical frequency attention mechanism is designed to help the model focus on local features and global features.
(3) An entropy maximum mean difference (EMMD) is designed. The entropy value is used to determine the bandwidth of the maximum mean difference Gaussian kernel, which can focus on the overall distribution of the samples.
(4) The diagnostic performance of IFEDAN under different loads and between different datasets is discussed, and the interpretability of the network is further analysed through visualization techniques. The results show that IFEDAN has good robustness and generalization performance.

The remaining part of the paper is organised as follows. Section 2 provides the introduction for the basic theory. Section 3 presents the proposed method. Section 4 presents the experimental validation and analysis. Section 5 is the conclusion.

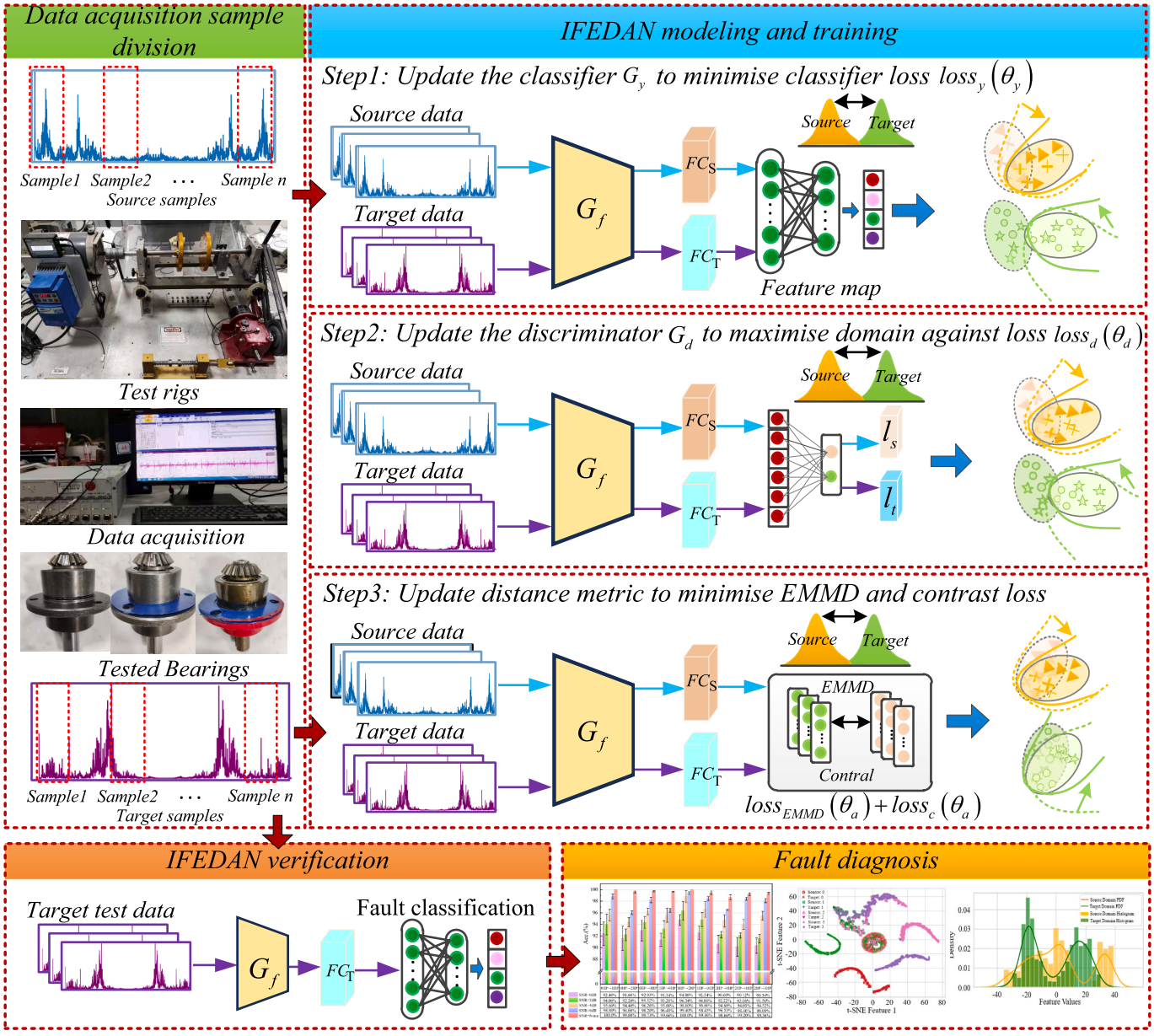


Fig. 1. Fault diagnosis flowchart of IFEDAN.

2. Methodologies

2.1. Problem definition

Domain adversarial neural network (DANN) is an adversarial trained transfer learning network [37], which consists of the feature extractor, domain discriminator and classifier. The feature extractor is shared by source and target domain samples to extract domain invariant features. In this paper, it is assumed that the condition Z_s is the source domain $X_s = \{x_i^s, y_i^s\}_{i=1}^{n_s}$, and the condition Z_t is the target domain $X_t = \{x_i^t\}_{i=1}^{n_t}$. Where n_s and n_t denote the number of samples in the source and target domains, respectively. The label space of the source and target domains is the same. However, the marginal probability distribution and conditional probability distribution are different for the source and target domains due to different working conditions. Therefore, the goal of cross-domain fault diagnosis is to train an intelligent model based on labelled source domain samples and unlabelled target domain samples.

2.2. Convolutional neural network (CNN)

2.2.1. Convolutional layer

CNN are commonly used to design feature extractors for DANN. CNN generally consist of the convolutional layer, pooling layer, batch normalization, activation function and fully connected layer [38]. Fault features are captured efficiently in the convolutional layer by using convolutional kernel.

2.2.2. Group normalization layer (GN)

Batch normalization (BN) is commonly used in CNN to normalise the inputs, thus reducing the bias of covariates internal to the model. However, BN is very sensitive to batch size. When the batch size is small, the model cannot compute the current mean and variance, and thus cannot estimate the covariate bias. To address the above problem, group normalization (GN) is used for covariate estimation.

2.2.3. Residual connection

Residual connection preserves feature information to reduce

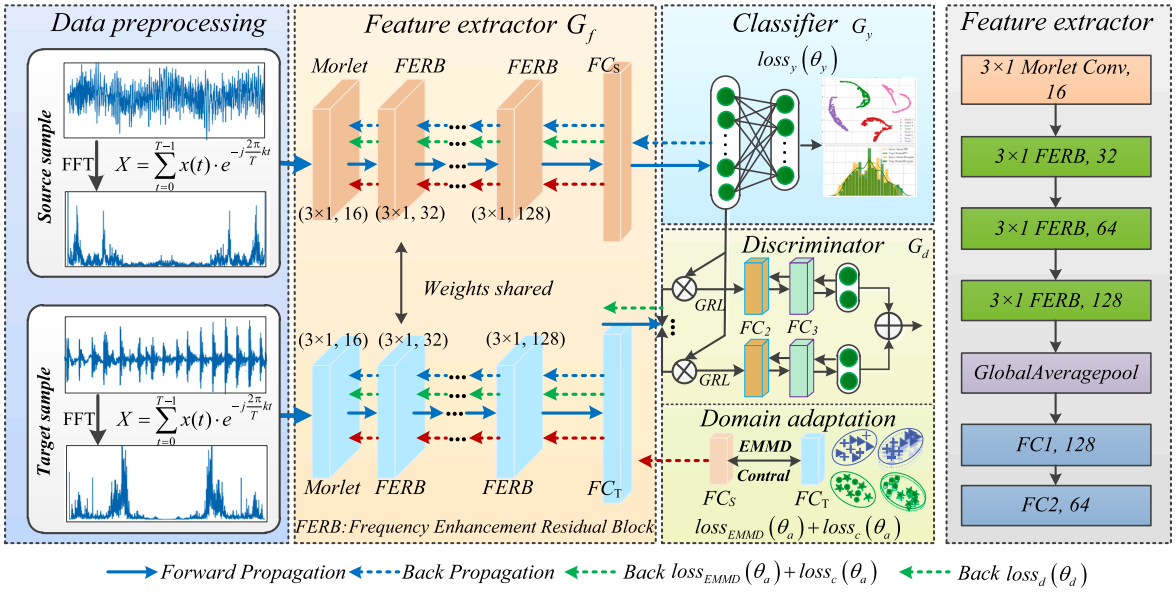


Fig. 2. The diagnostic flow of IFEDAN.

information missing during forward propagation [39]. In addition, it alleviates the problems of gradient disappearance and gradient explosion during model training.

3. Proposed method

An interpretable frequency-enhanced domain adaptive network (IFEDAN) is proposed to enhance the interpretability for cross-domain fault diagnosis. Fig. 1 shows the cross-domain fault diagnosis flowchart of IFEDAN, and its critical steps are summarised as follows:

Step1: Data acquisition and sample division. The vibration signals of bearings or gearboxes under different loads are collected. The data with different loads or different datasets are set as source and target domains, and FFT is performed. Then, divide the training set and test set of the source and target domains.

Step2: IFEDAN modelling and training. IFEDAN performs parameter updates. Specifically, first the classifier is updated with parameters and the classifier loss is minimised. Then, the domain discriminator is updated with parameters and the domain adversarial loss is maximised. Finally, the source and target domain features are updated with parameters and the EMMD and comparison loss is minimised.

Step3: IFEDAN Validation. Obtain the trained IFEDAN model, and the test set is input into the trained IFEDAN for fault diagnosis.

Step4: Fault diagnosis. The diagnostic results are obtained and analyzed.

The remainder of this section is organised as follows. First, we provide an overview of the overall framework of IFEDAN. Next, we describe the feature extractor of IFEDAN in detail. Finally, we describe the loss function and parameter update process of IFEDAN.

3.1. Description of IFEDAN

Many existing rolling bearing fault diagnosis methods are designed based on frequency domain [40], time domain features [41], or entropy [42]. These methods often directly input the extracted frequency domain and time domain features into neural networks, or use entropy values to measure the complexity of vibration signals. However, they overlook the importance of prior knowledge in the feature extraction process. The failure of rolling bearings produces periodic pulse characteristic, which is often drowned out by noise interference, making it difficult to identify. References [43,44] show that converting time-domain signals as frequency-domain signals can enhance periodic pulse characteristic and suppress noise interference. However, if only the frequency-domain signals are used as the model input without considering prior knowledge to guide feature extraction, the model cannot extract more domain-invariant features. Therefore, the proposed IFEDAN introduces prior knowledge (Morlet wavelet) for model weight initialization to extract time–frequency local features. In addition, a hierarchical frequency attention mechanism is designed to learn critical fault features. Fig. 2 shows the general framework of IFEDAN.

In Fig. 2, Fast Fourier Transform (FFT) is performed for the source and target domain samples. FFT can transform the time domain signal into the frequency domain to help the model better extract the frequencies of fault features. Then, a frequency-enhanced residual feature extractor is constructed. This feature extractor uses Morlet wavelet in the initial layer to enhance the interpretability. In addition, a frequency-enhanced residual block is designed, which hierarchically focuses on different low-frequency features and high-frequency features through the hierarchical frequency attention mechanism to mine more transferable features. Finally, domain adversarial training is implemented using classifier and discriminator, and the source and target domain test are fed into the trained network to achieve the classification.

3.2. Feature extractor of the model

3.2.1. Morlet convolution layer

CNN has powerful feature extraction capabilities. However, due to its 'black box' property, it lacks interpretability. Morlet is a wavelet function based on sine function modulation, which not only has time–frequency local characteristic, but also retains the amplitude and phase information for the signals. In order to enhance the interpretability for the model, a Morlet convolutional layer is constructed to replace the initialization layer of the feature extractor, which can perform time–frequency multi-resolution analysis on input signals in the early stages of model training. In addition, introducing Morlet wavelet into IFEDAN can improve the initialization of weight distributions and enhance the interpretability. The specific implementation of the Morlet convolution layer can be divided into two parts. First, the real and imaginary parts of the Morlet wavelet are defined as follows:

$$Mor_{real}(t,\sigma,f_c)=\frac{1}{\sigma}\exp(-\frac{t^2}{2\sigma^2})\cdot\cos(2\pi f_c t) \quad (1)$$

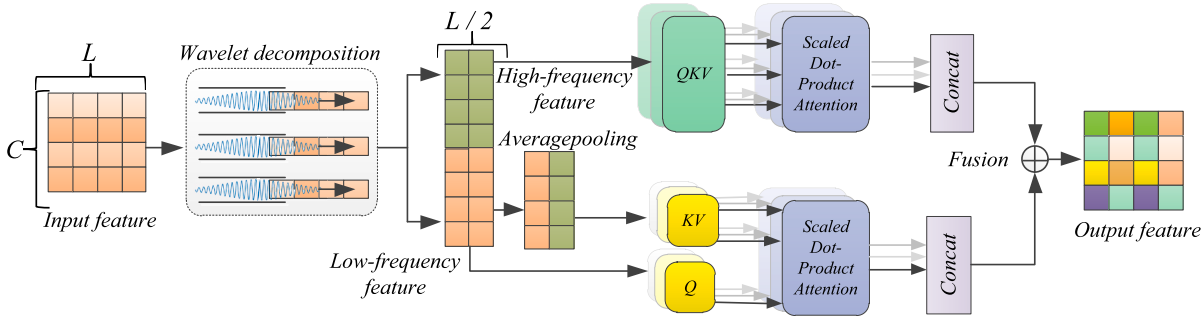


Fig. 3. Schematic diagram of the hierarchical frequency attention mechanism.

$$Mor_{imag}(t,\sigma,f_c)=\frac{1}{\sigma}\exp(-\frac{t^2}{2\sigma^2})\cdot\sin(2\pi f_c t) \tag{2}$$

where t is the time index, σ is the bandwidth parameter, and f_c is the center frequency. Then, the real and imaginary parts of the Morlet wavelet are used to perform the convolution operation, which are described as follows:

$$x_{real}(t)=F_{conv}(Mor_{real}(t,\sigma,f_c)*x(t)) \tag{3}$$

$$x_{imag}(t)=F_{conv}(Mor_{imag}(t,\sigma,f_c)*x(t)) \tag{4}$$

where $F_{conv}(*)$ is the convolution operation, $x_{real}(t)$ and $x_{imag}(t)$ are the real and imaginary parts of the convolution layer, respectively. $x(t)$ is the input signal. Finally, the real and imaginary parts of the convolution layer are computed to obtain the modulus of the wavelet convolution layer, which is described as follows:

$$Conv(t)=\sqrt{x_{real}^2(t)+x_{imag}^2(t)} \tag{5}$$

where $Conv(t)$ is the output of the wavelet convolution layer.

3.2.2. Hierarchical frequency attention mechanism

The faults of rotating machinery are usually presented in the periodic form. FFT converts the time-domain signals of rotating machinery into frequency-domain signals, making the frequency of faults more easily identifiable. However, the frequency resolution is susceptible to the length of the input samples. The length of the input samples is shorter, the model is more difficult to identify faults. Thus, in order to overcome the shortcomings of FFT, a hierarchical frequency attention mechanism is designed. This attention mechanism consists of discrete wavelet transform and self-attention mechanism. The hierarchical frequency attention mechanism is divided into two stages. In the first stage, the low-frequency features and high-frequency features of the input are obtained by discrete wavelet transform. In the second stage, the self-attention head is decomposed into high-frequency and low-frequency groups. The low frequency features and high frequency features are further obtained. The structure of the hierarchical frequency attention mechanism is shown in Fig. 3.

In Fig. 3, the hierarchical frequency attention mechanism is divided into three steps.

Step1: The input feature maps are discretised by discrete wavelet decomposition, and high-frequency features and low-frequency features are obtained, which are described as follows:

$$\begin{cases} X_H=DWT(x) \\ X_L=DWT(x) \end{cases}, x\in R^{C\times L},(X_L,X_H)\in R^{C\times L/2} \tag{6}$$

where x is the input feature map, $DWT(\cdot)$ is the discrete wavelet transform, and X_H and X_L are the high and low frequency features after wavelet decomposition.

Step2: the high-frequency features and low-frequency features are further processed using the self-attention mechanism. Specifically, the local window self-attention mechanism is used for high-frequency features to further mine the local high-frequency features. The global features are first obtained by average pooling for low-frequency features, and then the correlation between the low-frequency features and the global features is obtained using the query matrix, which are described as follows:

$$Att_H(x)=selfatt(X_H) \tag{7}$$

$$X_G=AP(X_L) \tag{8}$$

$$K=X_G\cdot W^K, Q=X_G\cdot W^Q, V=X_L\cdot W^V \tag{9}$$

$$Att_L(x)=softmax(\frac{QK^T}{\sqrt{d}})\cdot V \tag{10}$$

where $Att_H(x)$ is the high frequency feature, $selfatt(\cdot)$ is the window self-attention mechanism, X_G is the global feature obtained after average pooling, $AP(\cdot)$ is the average pooling operation, K is the key matrix, Q is the query matrix, s is the value matrix, and V is the low frequency feature.

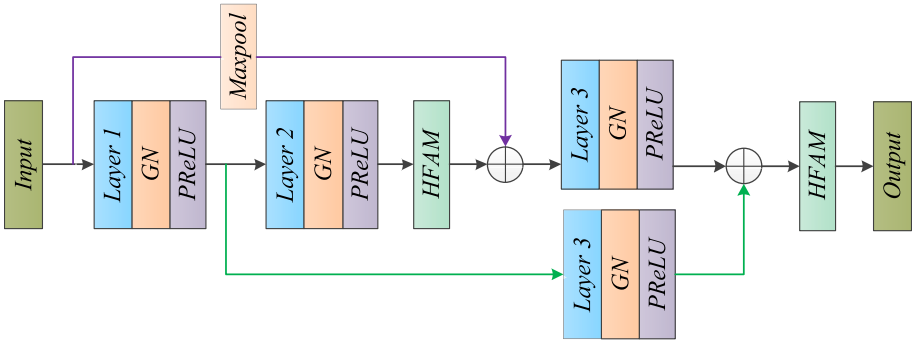


Fig. 4. Frequency-enhanced residual block.

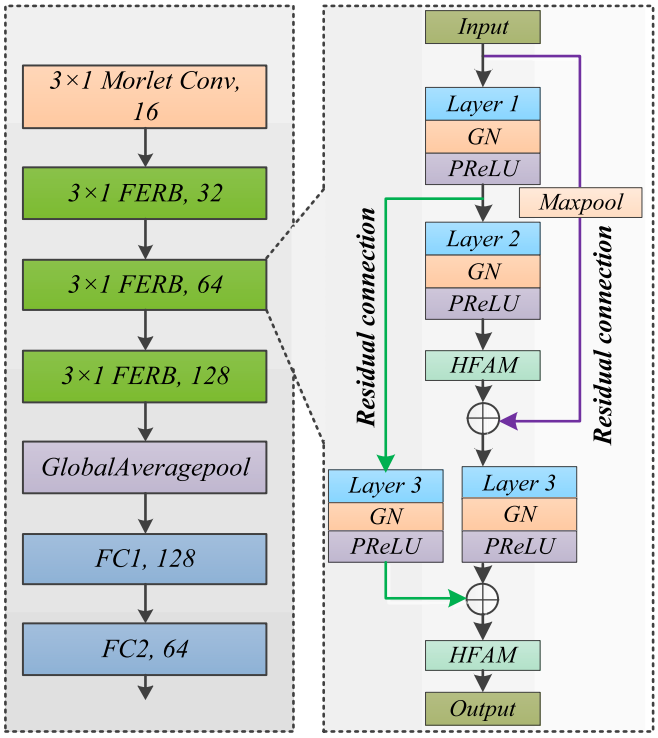


Fig. 5. Structure of feature extractor.

Step3: the high-frequency and low-frequency features obtained from each attention head are fused to obtain the output features, which are described as follows:

$$MSA_H = Concat(Att_H^1, Att_H^2, \ldots, Att_H^n)\cdot w \tag{11}$$

$$MSA_L = Concat(Att_L^1, Att_L^2, \ldots, Att_L^n)\cdot w \tag{12}$$

$$Output = MSA_H + MSA_L \tag{13}$$

where MSA_H is the high-frequency feature output of the multi-head attention mechanism and MSA_L is the low-frequency feature output of the multi-head attention mechanism. $Concat(\cdot)$ is the fusion operation, and $Output$ is the output feature map.

The hierarchical frequency attention mechanism can effectively capture the fault features in different frequency ranges. The high-frequency part focuses on local features and impulse signals, while the low-frequency part focuses on global features and periodic fault features. By fusing the high-frequency features with the low-frequency features, the input signals can be processed more comprehensively under different operating conditions and the accuracy of fault identification can be improved.

3.2.3. Frequency enhanced residual blocks

Residual networks can effectively solve the gradient vanishing problem and improve the training speed of the network [39]. However, in the cross-domain fault diagnosis task, its lack of screening ability for fault features leads to poor fault diagnosis performance. To solve this problem, this paper designs the frequency enhanced residual block, whose structure is shown in Fig. 4.

The frequency enhanced residual block has the following differences from the traditional residual block:

(1) Use group normalization (GN) to replace the batch normalization layer. the GN layer can be used in small batch training scenarios to improve the robustness.
(2) Use PReLU to replace ReLU. PReLU can learn features in negative regions and enhance the feature learning capability.
(3) Embed the frequency attention mechanism into the residual block. The frequency attention mechanism can not only focus on local features and impulse signals, but also pay attention to global features and periodic fault features.
(4) A residual branch is added to the traditional residual block, which avoids the missing for fault features and helps the model to mine more domain-invariant features.

3.2.4. Frequency enhanced residual feature extractor

A frequency enhanced residual feature extractor is designed using the Morlet convolution layer and frequency enhanced residual blocks, as shown in Fig. 5. The frequency enhanced residual feature extractor consists of a Morlet convolution layer, three frequency enhanced residual blocks, an average pooling layer and a fully connected layer. Each frequency enhanced residual block is passed between them using skip connection. The source domain samples and the target domain samples share weights in the frequency enhanced residual feature extractor. The frequency enhanced residual feature extractor can adequately extract low-frequency and high-frequency fault features to more comprehensively deal with input signals and improve the accuracy of cross-domain fault identification.

3.3. Domain adaptive loss function

3.3.1. EMMD loss function

In transfer learning, maximum mean difference is widely used to measure the distribution distance between two domains. However, the performance of maximum mean difference is affected by the Gaussian kernel. Currently, the Euclidean distance is usually used to determine the Gaussian kernel. The Euclidean distance also changes with model training, which causes the MMD to become unstable. To solve the above problem, this paper uses the entropy value between two domains to select the bandwidth for the Gaussian kernel. The entropy value can focus on the overall distribution of the samples, thus solving the problem that the MMD becomes unstable during the training process. Specifically, the Gaussian kernel function is described as follows:

$$K(\psi_s, \psi_t) = \exp\left(\frac{\left\|\psi_s(x_i) - \psi_t(x_j)\right\|^2}{\xi+\varepsilon}\right) \tag{14}$$

where $Kernel(\cdot)$ is the Gaussian kernel function, ψ_s is the source domain, ψ_t is the target domain, ξ is the bandwidth of the Gaussian kernel function, and ε is a constant for preventing errors. Then, the bandwidth of the Gaussian kernel function is calculated using the entropy value, which is described as follows:

$$\begin{aligned} \xi &= |\psi_s - \psi_t| \\ &= -\frac{1}{d}\left|\sum_{i=1}^{d}\log(\sigma_i+\varepsilon) - \sum_{i=1}^{d}\log(\sigma_j+\varepsilon)\right| \end{aligned} \tag{15}$$

where d is the total number of features, σ_i is the source domain sample standard deviation $\sigma_i = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_i)^2}$. σ_j is the source domain sample standard deviation $\sigma_j = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_j-\mu_j)^2}$.

Finally, the entropy maximum mean difference (EMMD) can be described as follows:

$$EMMD(\psi_s,\psi_t)=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq 1}^{n}K(\psi_s(x_i),\psi_s(x_i))+\frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq 1}^{m}K(\psi_t(x_j),\psi_t(x_j))-\frac{1}{nm}\sum_{i=1}^{n}\sum_{j\neq 1}^{m}K(\psi_s(x_i),\psi_t(x_j)) \tag{16}$$

where n is the total number of samples in the source domain and m is the total number of samples in the target domain.

3.3.2. Contrast loss

Contrast loss can help the same samples to be closer and the different sample to be farther away in the feature space. The contrast loss can effectively learn the similarity and difference between the samples, so as to improve the performance for the cross-domain fault diagnosis [45]. The contrast loss is described as follows:

$$loss_c=\frac{1}{N}\sum_{i=1}^{N}\left[\|\psi_s(x_i)-\psi_t(x_i)\|^2+\max(0,\tau-\|\psi_s(x_i)-\psi_t(x_j)\|^2)\right] \tag{17}$$

where N is the batch size and τ is the boundary value.

3.4. Overall objective and optimization

The overall optimization objective of IFEDAN can be summarized as follows: minimize the loss of the domain classifier, the EMMD loss and the contrast loss, and maximize the loss of the domain discriminator. Therefore, the overall loss of IFEDAN can be described as follows:

$$L(\theta_y,\theta_a,\theta_d)=\min(loss_y(\theta_y))+\min_a(loss_{EMMD}(\theta_a)+loss_c(\theta_a))+\max_d(loss_d(\theta_d)) \tag{18}$$

where θ_y, θ_a and θ_d are domain classifier, domain adaptive, and domain discriminator parameters, respectively. $loss_y(\cdot)$ is the domain classifier loss, $loss_{EMMD}(\cdot)$ is the EMMD loss, $loss_c(\cdot)$ is the contrast loss, and $loss_d(\cdot)$ is the domain discriminator loss.

In IFEDAN, we introduce a gradient reversal layer (GRL) to further help the feature extractor to learn more domain invariant features, which are described as follows:

$$\left(\widehat{\theta}_y,\widehat{\theta}_a,\widehat{\theta}_d\right)=\text{argmin}L\left(\theta_y,\theta_a,\widehat{\theta}_d\right) \tag{19}$$

$$\widehat{\theta}_d=\text{argmin}L\left(\widehat{\theta}_y,\widehat{\theta}_a,\theta_d\right) \tag{20}$$

where $\widehat{\theta}_y$, $\widehat{\theta}_a$ and $\widehat{\theta}_d$ are the estimates of the parameters. The pseudo-code flowchart for IFEDAN is shown in Algorithm 1.

Algorithm 1 Training and testing procedure for IFEDAN

Input: Source domain labelled data $X_s=\{x_i^s,y_i^s\}_{i=1}^{n_s}$, target domain unlabelled data $X_t=\{x_i^t\}_{i=1}^{n_t}$

1: Set hyperparameters, such as: convolution layer, pooling layer, learning rate, batch size
2: Divide the training set and test set for source domain and target domain
3: Initialise IFEDAN weights and biases
4: **For** each training epoch **do**
5: If IF = Train then
6: Perform the following operations on each epoch
7: Calculate the output of the classifier
8: Calculate the loss of the classifier $loss_y(\cdot)$
9: Calculate the outputs of the discriminator and EMMD
10: Calculate the loss of the discriminator $loss_d(\cdot)$
11: Calculate $loss_{EMMD}(\cdot)$ and $loss_c(\cdot)$ according to Eqs. (21) and (22)
12: Obtain the optimised objective function $L(\theta_y,\theta_a,\theta_d)$ according to Eq. (23)

(continued on next column)

(continued)

Algorithm 1 Training and testing procedure for IFEDAN

13: Update model parameters
14: **End for**
15: Save the trained model

Output: Test data is loaded into the trained model, and the diagnostic results are obtained for the test samples.

4. Experimentation and analysis

In this section, one gearbox dataset and two bearing datasets are used for experimental validation and analysis. The model is trained on a Windows 10 system configured with an AMD Ryzen 5-4600H processor and 16 GB RAM. The experimental environment is Python 3.6 and TensorFlow 2.0.0.

4.1. Dataset Description and data processing

4.1.1. Dataset Description

(1) The roller gear (RG) failure simulation platform consists of an AC asynchronous motor, a stand gearbox, a variable frequency controller, and a loading reset. The test platform is shown in Fig. 6. The tested gear of this test platform is a bevel gear. The gear state is classed into tooth wear, full tooth breakage, half tooth breakage and healthy state. A total of 0HP, 1HP, 2HP and 3HP data under 20 Hz and 30 Hz are collected through the frequency controller and loading device. The data of four working conditions at 30 Hz are selected in this paper. Table 1 contains the information of RG dataset.
(2) Fig. 7 shows the bearing test platform of Lanzhou University of Technology (LUT). The platform consists of a three-phase AC

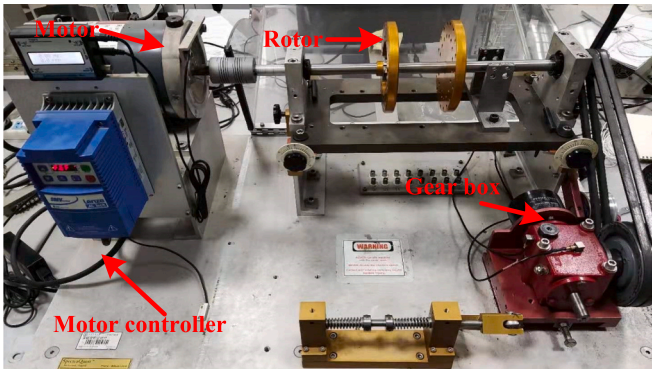


Fig. 6. The diagram of the RG test rig.

Table 1
Division of the dataset.

Dataset	Condition	Fault mode	Label
RG(A)	0HP (A_0)	Health state (HS)	0
	1HP (A_1)	Complete tooth breakage (CT)	1
	2HP (A_2)	Half tooth breakage (HT)	2
	3HP (A_3)	Tooth wear (TW)	3
LUT(B)	1449 r/min (0HP/B_0)	Rolling body fault (RB)	0
	1378 r/min (1HP/B_1)	Inner ring fault (II)	1
	1251 r/min (2HP/B_2)	Outer ring fault (OI)	2
	1130 r/min (3HP/B_3)	Mixed fault (MF)	3
CWRU(C)	1790 r/min (0HP/C_0)	Roller fault (RF)	0
	1772 r/min (1HP/C_1)	Inner ring fault (IF)	1
	1750 r/min (2HP/C_2)	Outer ring fault (OF)	2
	1730 r/min (3HP/C_3)	Normal state (NS)	3

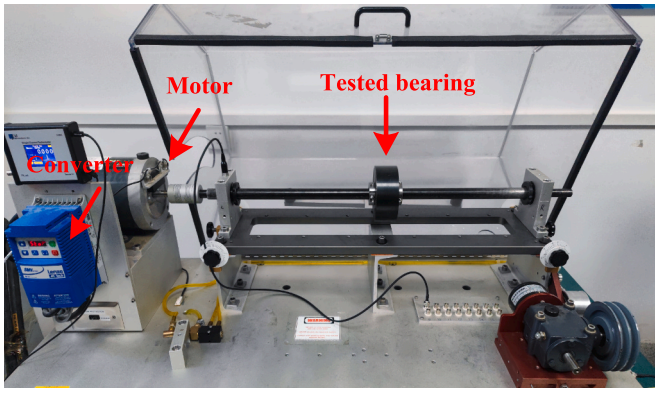


Fig. 7. The diagram of the RG test rig.

motor, a drive shaft and a tested bearing. In this study, bearing vibration signals are collected at four speeds (1130, 1251, 1378 and 1449 rpm). The bearing states are rolling body fault, inner ring fault, outer ring fault and combined fault. Table 1 contains information aout the LUT dataset.

(3) The Case Western Reserve University (CWRU) bearing dataset is widely used for validation. In this study, vibration signals of 1730r/min, 1750r/min, 1772r/min and 1790r/min at 12KHz are used. The bearing states are categorized as roller fault, inner ring fault, outer ring fault and normal state.

Table 1 contains information about the above dataset. The total number of samples in the source and target domains is 600. The training set and test set are divided according to the ratio of 4:1.

4.1.2. *Data processing*

FFT is used for data processing. FFT can transform the time domain

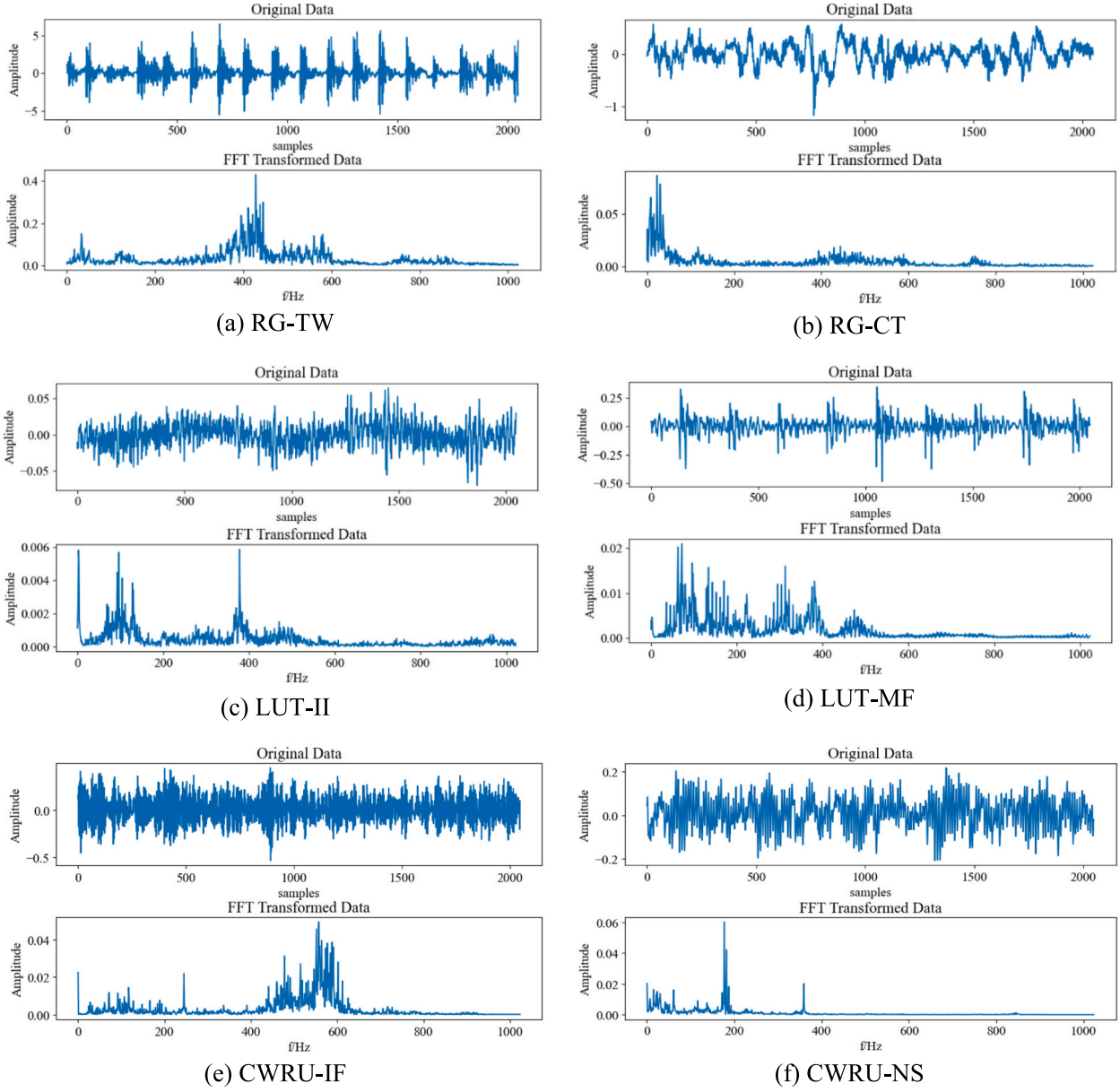


Fig. 8. Visualization of vibration signals for different datasets.

Table 2
Model parameters of IFEDAN.

Names	Layers	Operators	Stride/Size	Output
Feature extractor		Morlet Conv	2/3 × 1	(1024,16)
	FERB	Conv (Layer)	2/3 × 1	(512,32)
		GN/PReLU	–	
		HFAM	–	
	FERB	Conv (Layer)	2/3 × 1	(256,64)
		GN/PReLU	–	
		HFAM	–	
	FERB	Conv (Layer)	2/3 × 1	(128,128)
		GN/PReLU	–	
		HFAM	–	
Discriminator	FC1	Dense (ReLU)	–	128
	FC2	Dense (ReLU)	–	64
		Sigmoid	–	1
Classifier	FC3	Dense (ReLU)	–	128
	FC4	Dense (ReLU)	–	64
		Softmax	–	Class

signals into frequency domain, which can easily identify the frequency information of the fault features. The equation of FFT is described as follows:

$$X = \sum_{t=0}^{T-1} x(t) \cdot e^{-j\frac{2\pi}{T}kt} \tag{21}$$

where $x(t)$ is the input signal, T is the total length of the input signals, and k is the frequency index. Fig. 8 shows the visualization results for the different datasets.

4.2. Model parameters and comparison methods

4.2.1. Model parameters

Table 2 shows the network parameters of IFEDAN. The input sample length is 1024. Adam is selected as the optimization algorithm for model training, and the hyperparameters of the model are determined through grid search. The learning rate is 0.001, and the batch size is 64. In addition, to demonstrate the reliability and stability of IFEDAN, the experimental results are the average of 10 experiments.

4.2.2. Comparison methods

IFEDAN is compared with eight other state-of-the-art cross-domain methods, including DDTLN [46], TFN [47], WIDAN [48], DADDN [49], MMCLE [50], FJDMA [51], WKWJDAN [52], and DFD [53]. DDTLN is a joint distribution adaptive method that reduces domain differences through maximum mean difference and coral distribution. TFN is a new time–frequency network that uses kernel functions to design time–frequency convolutions and extract time–frequency information. WIDAN is a wavelet domain adaptive network using physical information to achieve cross-domain machine fault diagnosis. DADDN reduces negative transfer of target domain features through a dual-domain adversarial mechanism and weighted joint adaptive distribution. MMCLE is a multi-channel and multi-scale domain adversarial neural network that extracts fault features through multi-scale and LSTM. FJDMA is a domain adversarial model that performs the feature and joint distribution difference alignment, achieving global alignment through an attention mechanism. WKWJDAN is a multi-kernel weighted joint domain adaptive network that enhances domain confusion by utilising multi-kernel conditional maximum difference. DFD is a neural network based on information theory, which extracts domain-related knowledge through contrastive learning.

The above comparative methods perform cross-domain fault diagnosis under variable operating conditions or between different datasets. In addition, the above comparative methods perform domain difference metrics and attention mechanisms. Therefore, the diagnostic performance of IFEDAN can be adequately compared.

4.3. Cross-domain diagnostic results under variable loads

4.3.1. RG fault diagnosis results

In real industrial scenarios, the vibration signals of gearboxes usually are varied with the load, resulting in the differences of data distribution for the vibration signals. In addition, noisy environments cause the feature information to become more complex, further increasing the data distribution discrepancies. Thus, it is of great significance to investigate the cross-domain diagnosis of gearboxes under variable loads and noisy environments. In this paper, Gaussian white noise with signal-to-noise ratio (SNR) of 0 dB, 2 dB, 4 dB, and 6 dB is added to the vibration signals for IFEDAN validation. The results of the fault diagnostics under different noise environments are shown in Fig. 9.

In Fig. 9, the average diagnostic accuracy of IFEDAN is 99.51 % when

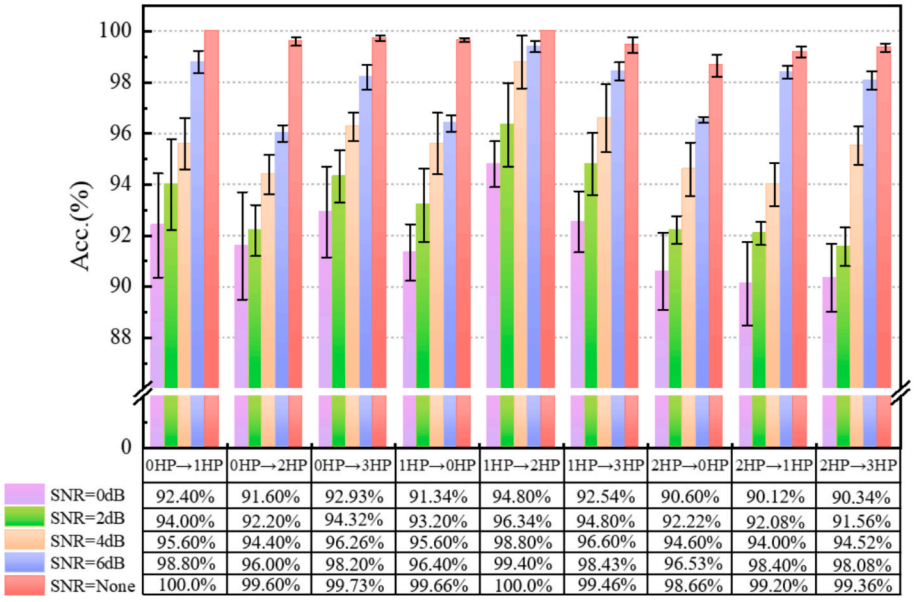
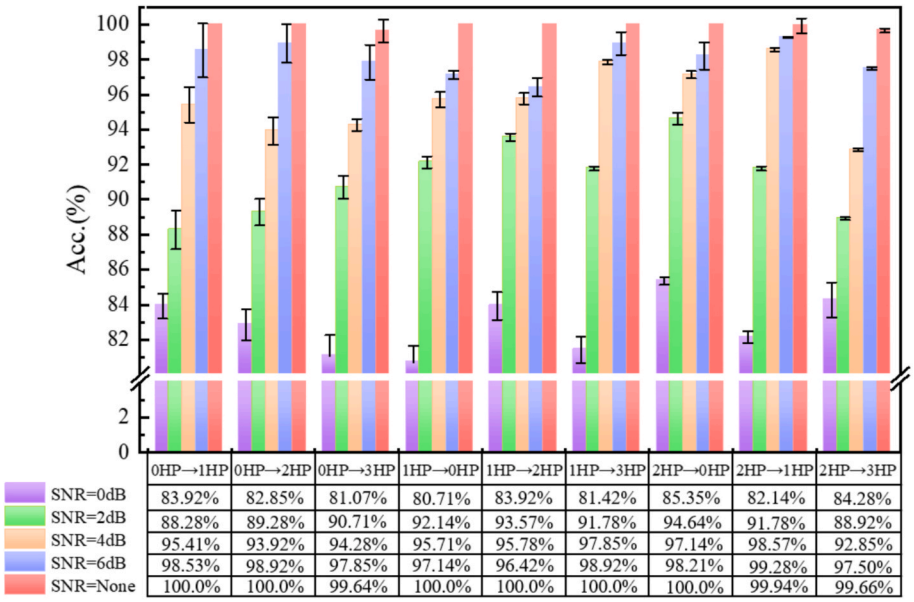


Fig. 9. Fault diagnosis results of RG.



	0HP→1HP	0HP→2HP	0HP→3HP	1HP→0HP	1HP→2HP	1HP→3HP	2HP→0HP	2HP→1HP	2HP→3HP
SNR=0dB	83.92%	82.85%	81.07%	80.71%	83.92%	81.42%	85.35%	82.14%	84.28%
SNR=2dB	88.28%	89.28%	90.71%	92.14%	93.57%	91.78%	94.64%	91.78%	88.92%
SNR=4dB	95.41%	93.92%	94.28%	95.71%	95.78%	97.85%	97.14%	98.57%	92.85%
SNR=6dB	98.53%	98.92%	97.85%	97.14%	96.42%	98.92%	98.21%	99.28%	97.50%
SNR=None	100.0%	100.0%	99.64%	100.0%	100.0%	100.0%	100.0%	99.94%	99.66%

Fig. 10. Fault diagnosis results of LUT.

Table 3
Cross-domain diagnostic results between RG and LUT (%).

Methods	$A_0 \rightarrow B_0$	$A_1 \rightarrow B_1$	$A_2 \rightarrow B_2$	$A_3 \rightarrow B_3$	$B_0 \rightarrow A_0$	$B_1 \rightarrow A_1$	$B_2 \rightarrow A_2$	$B_3 \rightarrow A_3$
IFEDAN	88.23 ± 0.33	89.85 ± 0.76	94.70 ± 2.30	92.20 ± 0.94	98.07 ± 0.33	95.14 ± 0.88	95.88 ± 0.23	96.41 ± 0.44
DDTLN	85.83 ± 1.02	86.47 ± 0.44	80.58 ± 0.77	77.53 ± 1.22	84.72 ± 1.22	86.87 ± 0.22	83.91 ± 0.62	83.82 ± 0.88
TFN	81.04 ± 0.12	84.58 ± 0.98	82.50 ± 0.66	86.04 ± 0.42	92.27 ± 0.55	84.58 ± 1.05	91.66 ± 0.45	90.83 ± 0.66
WIDAN	82.92 ± 0.67	85.00 ± 0.17	91.67 ± 0.07	84.16 ± 0.54	91.45 ± 0.05	86.56 ± 1.22	89.79 ± 0.77	93.54 ± 0.28
DADDN	77.50 ± 0.74	83.75 ± 0.25	80.41 ± 0.23	84.58 ± 0.96	68.33 ± 0.44	80.09 ± 0.62	72.08 ± 0.13	75.00 ± 1.02
MMCLE	82.50 ± 0.77	84.37 ± 0.33	81.83 ± 0.84	85.05 ± 0.21	86.91 ± 0.32	85.62 ± 0.76	82.70 ± 1.01	83.75 ± 0.14
FJDMA	80.44 ± 0.16	82.04 ± 1.02	89.67 ± 0.48	87.08 ± 0.50	93.33 ± 0.72	92.50 ± 0.06	91.33 ± 0.33	93.16 ± 0.18
WKWJDAN	76.50 ± 0.55	76.67 ± 0.08	86.83 ± 0.44	80.83 ± 0.67	72.66 ± 0.72	81.67 ± 0.34	75.33 ± 0.84	83.33 ± 0.07
DFD	74.35 ± 0.72	72.08 ± 1.06	76.00 ± 0.50	78.66 ± 0.16	70.37 ± 0.22	80.44 ± 0.21	77.50 ± 0.13	82.83 ± 0.20

SNR=None. When SNR = 6 dB, 4 dB, 2 dB and 0 dB, the average diagnostic accuracy of IFEDAN is 97.80 %, 95.70 %, 93.41 % and 91.85 %, respectively. This indicates that the diagnostic accuracy of IFEDAN is reduced as the SNR decreases. This indicates that the noisy environment makes the feature information complex, which affects the feature extraction performance. However, the diagnostic accuracies of IFEDAN are all above 90.0 %, which further indicates that IFEDAN has good stability and robustness.

4.3.2. *LUT fault diagnosis results*

To verify its generalization performance, the LUT bearing dataset is chosen for the study. The experimental parameters and noise environment are set the same as in the previous section. The experimental results are shown in Fig. 10. In Fig. 10, when SNR=None, the average diagnosis accuracy of IFEDAN is 99.91 %. This indicates that IFEDAN can accurately identify bearing faults under variable loads and non-noise environments. When SNR = 6 dB, 4 dB, 2 dB and 0 dB, the average diagnostic accuracy of IFEDAN is 98.08 %, 95.72 %, 91.23 % and 82.85 %, respectively. This indicates that the frequency enhanced residual block of IFEDAN can extract more domain-invariant features and has good generalization performance in strong noise environments.

4.4. *Cross-domain diagnostic results with different datasets*

4.4.1. *Cross-domain diagnostic results between RG and LUT*

Intelligent diagnostic models not only need to be competent in fault diagnosis tasks for variable loads, but also need to satisfy cross-domain diagnostic tasks between different mechanical devices. Therefore, we conduct cross-domain diagnosis studies between different datasets. The diagnosis results are shown in Table 3.

In Table 3, IFEDAN shows better cross-domain diagnosis results, and its average diagnosis accuracy is 93.81 %. Compared with the other eight methods, the diagnosis accuracy of IFEDAN is improved by 10.1 %, 7.13 %, 5.68 %, 16.1 %, 9.72 %, 5.12 %, 14.58 %, and 17.28 %, respectively. At the same time, the diagnostic results of TFN, WIDAN, and FJDMA were significantly better than those of DADDN. This shows that IFEDAN can help the model to extract more domain-invariant features to enhance the fault identification accuracy by designing a better feature extractor. Secondly, we find that the diagnosis accuracy when using LUT as the source domain is significantly higher than that when using RG as the source domain. For example, the diagnostic accuracy of $B_0 \rightarrow A_0$ is higher than that of $A_0 \rightarrow B_0$. This indicates that different datasets contain different fault information, and setting the dataset with rich fault information as the source domain helps to improve the cross-domain diagnostic accuracy.

In addition, in order to quantify the cross-domain diagnosis performance for different methods, we visualise the confusion matrix for $B_0 \rightarrow A_0$. The results are shown in Fig. 11. In Fig. 11, IFEDAN, DDTLN, TFN, WIDAN, DADDN, MMCLE and DFD have high identification accuracies for the health state (label 0). For tooth wear, complete tooth breakage and half tooth breakage, the identification accuracy of the comparison methods is poor. This is due to that the features are similar for tooth face wear, complete tooth breakage and half tooth breakage. These methods cannot mine more transferable features, resulting in their poor cross-domain diagnostic performance. WKWJDAN has a high misclassification rate for health state (label 0) and complete tooth breakage (label 2).

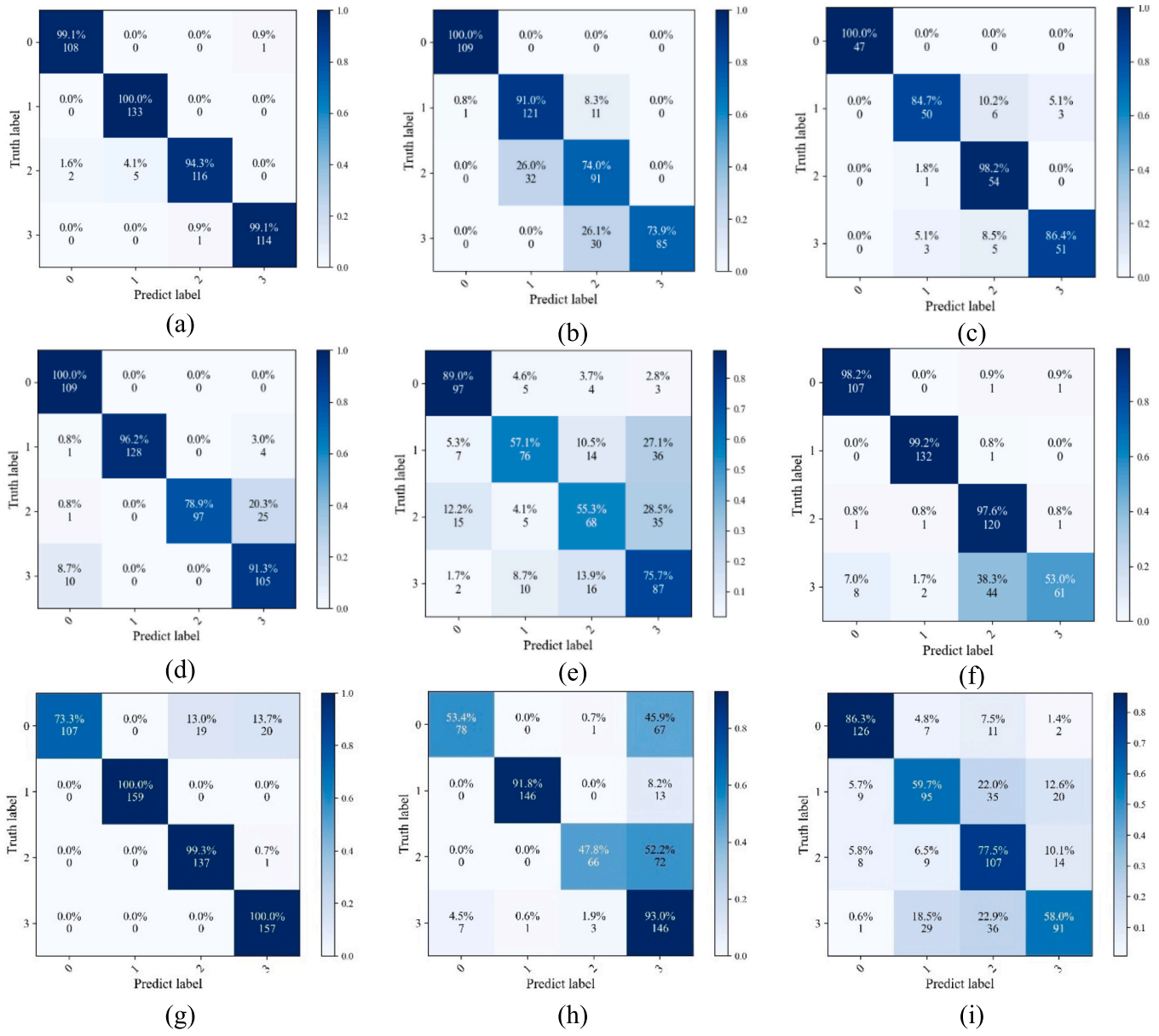


Fig. 11. Cross-domain diagnosis results of different methods. (a) IFEDAN, (b) DDTLN, (c) TFN, (d) WIDAN, (e) DADDN, (f) MMCLE, (g) FJDMA, (h) WKWJDAN, (i) DFD.

Table 4
Cross-domain diagnostic results between LUT and CWRU (%).

Methods	$B_0 \rightarrow C_0$	$B_1 \rightarrow C_1$	$B_2 \rightarrow C_2$	$B_3 \rightarrow C_3$	$C_0 \rightarrow B_0$	$C_1 \rightarrow B_1$	$C_2 \rightarrow B_2$	$C_3 \rightarrow B_3$
IFEDAN	90.62 ± 0.42	92.02 ± 0.15	94.37 ± 0.33	98.12 ± 0.42	97.08 ± 0.22	99.16 ± 0.08	99.01 ± 0.03	98.54 ± 0.22
DDTLN	80.09 ± 0.74	84.27 ± 0.24	90.41 ± 0.13	85.58 ± 0.21	92.94 ± 1.36	94.26 ± 0.33	93.23 ± 0.64	92.50 ± 1.02
TFN	81.47 ± 1.54	82.50 ± 0.74	82.20 ± 1.22	86.91 ± 0.57	91.17 ± 0.44	96.47 ± 0.14	86.91 ± 1.12	87.64 ± 0.96
WIDAN	85.00 ± 1.05	89.00 ± 0.50	88.64 ± 0.50	92.50 ± 0.05	94.50 ± 0.05	91.50 ± 0.25	93.50 ± 0.05	94.00 ± 0.55
DADDN	81.90 ± 0.50	83.33 ± 1.05	82.64 ± 0.50	86.12 ± 0.25	83.00 ± 0.55	90.07 ± 1.55	90.00 ± 0.55	87.50 ± 0.25
MMCLE	84.37 ± 0.34	83.41 ± 0.44	84.12 ± 0.17	88.33 ± 0.24	93.12 ± 0.12	85.20 ± 0.83	91.58 ± 0.22	91.04 ± 0.66
FJDMA	86.04 ± 0.18	82.96 ± 0.72	91.00 ± 0.11	90.67 ± 0.62	92.36 ± 0.24	95.50 ± 0.22	92.83 ± 0.13	94.16 ± 0.50
WKWJDAN	82.27 ± 0.34	79.93 ± 0.48	81.80 ± 0.04	83.33 ± 0.26	90.67 ± 0.08	89.53 ± 0.27	90.02 ± 0.68	86.67 ± 0.05
DFD	71.50 ± 1.22	73.33 ± 0.08	72.84 ± 1.06	75.46 ± 0.58	84.36 ± 0.66	85.83 ± 0.74	82.16 ± 0.33	84.02 ± 0.66

This is because WKWJDAN lacks the guidance from the attention mechanism, making it difficult to capture critical fault features. However, the frequency enhanced residual block in IFEDAN can effectively capture the fault features in different frequency ranges. The high-frequency part focuses on local features and impulse signals, while the low-frequency part focuses on global features and periodic fault features. By fusing the high-frequency features with the low-frequency features, the robustness can be improved for cross-domain fault diagnosis.

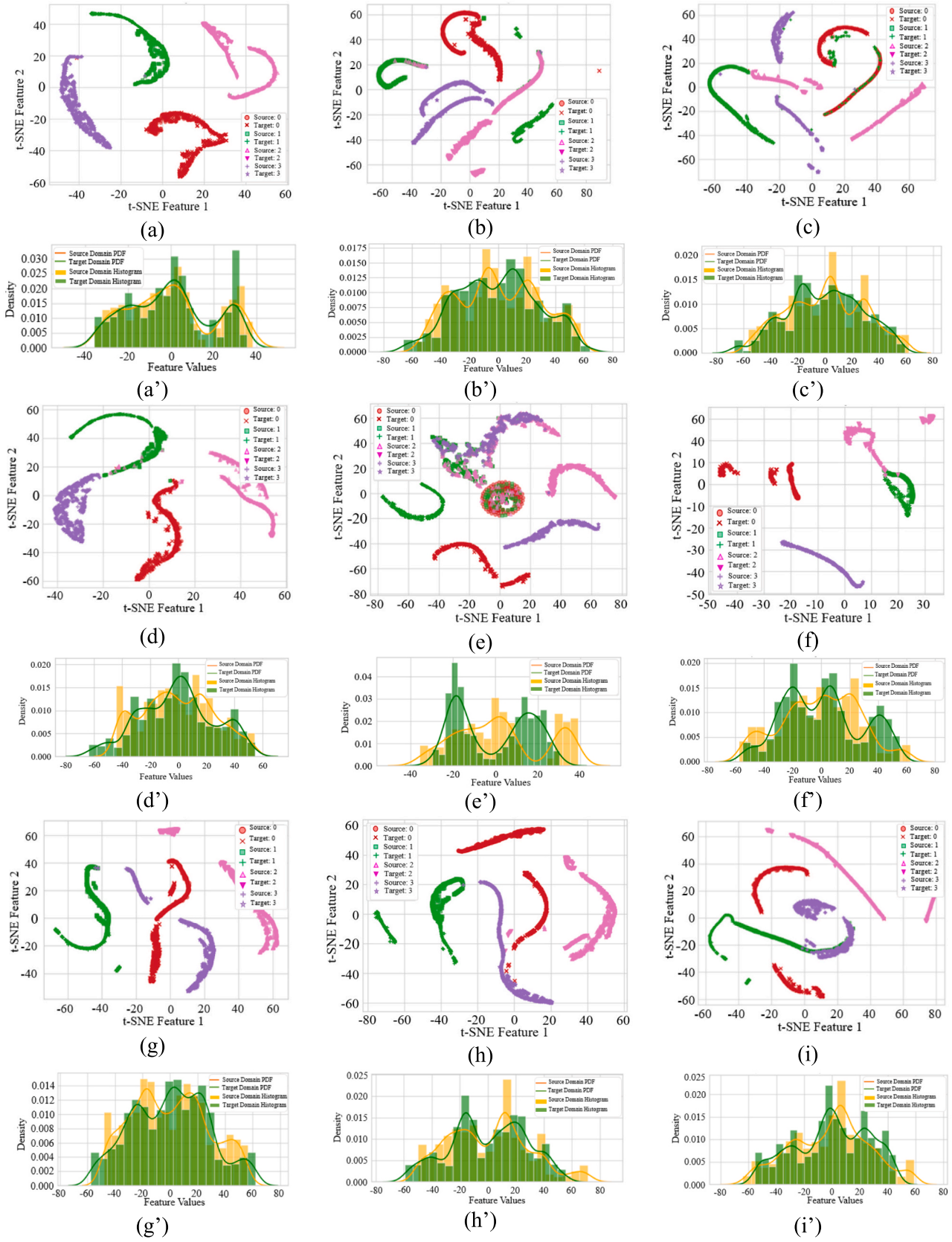


Fig. 12. Cross-domain diagnosis results of different methods. (a) IFEDAN, (b) DDTLN, (c) TFN, (d) WIDAN, (e) DADDN, (f) MMCLE, (g) FJDMA, (h) WKWJDAN, (i) DFD, (a′) IFEDAN, (b′) DDTLN, (c′) TFN, (d′) WIDAN, (e′) DADDN, (f′) MMCLE, (g′) FJDMA, (h′) WKWJDAN, (i′) DFD.

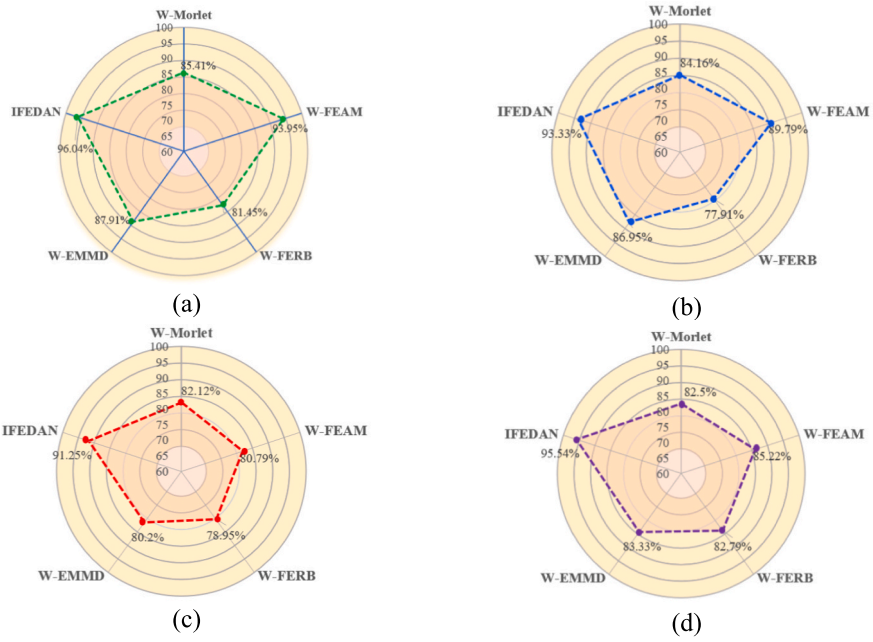


Fig. 13. Diagnostic results of ablation experiments. (a) $A_0 \rightarrow C_0$, (b) $A_1 \rightarrow C_1$, (c) $A_2 \rightarrow C_2$, (d) $A_3 \rightarrow C_3$.

4.4.2. Cross-domain diagnosis results between LUT and CWRU

In order to further verify the generalization performance of IFEDAN, we perform cross-domain diagnosis study between CWRU and LUT. The results are shown in Table 4. In Table 4, the diagnosis accuracy of IFEDAN is higher than the comparison methods, and its average diagnosis accuracy is 96.11 %. Compared with the eight comparison methods, the diagnosis accuracy of IFEDAN is improved by 6.95 %, 9.21 %, 5.03 %, 10.54 %, 8.47 %, 5.42 %, 10.58 %, and 17.42 %, respectively. Meanwhile, IFEDAN shows more stable diagnostic performance in all the transfer learning tasks, and the diagnostic accuracies are above 90 %. The results illustrate that IFEDAN has better generalization performance in cross-domain diagnosis. Compared with TFN and WIDAN that introduce physical knowledge, IFEDAN can further reduce the domain discrepancy by entropy maximum mean difference. Compared with DDTIL, DADDN and MMCLE, IFEDAN helps to improve the fault diagnosis accuracy for the model by constructing better feature extractor. Compared with WKWJDAN and DFD, IFEDAN can further focus on critical features and improve the fault diagnosis accuracy of the model. Compared with FJDMA, the wavelet convolution layer can enhance the model's ability to capture fault features.

In order to further visualise the domain adaptive capability of IFEDAN, we perform t-SNE and probability density profile visualization for $C_0 \rightarrow B_0$. The results are shown in Fig. 12. T-SNE can map the high-dimensional features to the two-dimensional space to better observe the feature extraction capability for the model. The probability density profile can show how closely the source and target domains match. In Fig. 12(a), IFEDAN can correctly identify the fault classes of different domains, and the intra-class distance between the source and target domains is smaller and the inter-class distance is larger. In Fig. 12(a′), we can also see that there is more overlap between the source domain (orange) and the target domain (green). In Fig. 12(b)–(i), although the eight methods perform preliminary classification of faults classes across different domains, however, the high misclassification rates in different classes lead to poor fault diagnosis accuracy. Additionally, the probability density curves in Fig. 12(b′)–(i′) also reveal poor domain matching between the source domain and the target domain. This further indicates that cross-domain diagnosis between different datasets is a challenging task.

4.5. Ablation experiments and interpretable analyses

4.5.1. Ablation experiments

To verify the effect of different components for IFEDAN, we selected $A_0 \rightarrow C_0$, $A_1 \rightarrow C_1$, $A_2 \rightarrow C_2$ and $A_3 \rightarrow C_3$ for the ablation experiments. The comparison methods are W-Morlet (without Morlet layer), W-FEAM (without frequency enhanced attention mechanism), W-FERB (without frequency enhanced residual block), and W-EMMD (without entropy maximum mean difference). The parameter settings of the comparison methods are the same as IFEDAN. The experimental results are shown in Fig. 13. First, the frequency enhanced residual block has the greatest impact on the performance for IFEDAN by comparing Fig. 13(a)–(d). In Fig. 13(a), the diagnostic accuracy of W-FEAM is 14.59 % lower than that of IFEDAN. This indicates that the frequency enhanced residual block can focus on critical high-frequency and low-frequency fault features and mine more domain-invariant features to extract the diagnostic accuracy for IFEDAN. In Fig. 13(b) and (c), the diagnostic accuracy of IFEDAN is improved by 9.17 % and 6.38 % compared to W-Morlet and W-EMMD, respectively. This indicates that the domain differences of the samples can be effectively captured by adding physical knowledge (Morlet wavelet and entropy value).

4.5.2. Interpretability analysis

In order to illustrate the post hoc interpretability for IFEDAN, we select the samples of different fault states for RG, LUT and CWRU to perform the visualization analysis. The class activation mapping (CAM) is used to extract the Morlet convolution layer weight mapping map and weight values. The results are shown in Fig. 14. In Fig. 14, the Morlet convolution layer can assign different weight values for regions in different frequency bands, thus focusing on more important fault features. Specifically, in Fig. 14(a) and (b), regions with frequencies between 300 and 600 are assigned higher weights. In Fig. 14 (c) and (d), regions with frequencies between 400 and 700 are given higher weights. In Fig. 14(e) and (f), the frequency of 0–50 is given higher weight. The above results show that the Morlet convolution layer can assign different weights for different samples in different frequency regions. Higher weights are assigned to critical features by the Morlet convolution layer to improve the cross-domain diagnostic performance.

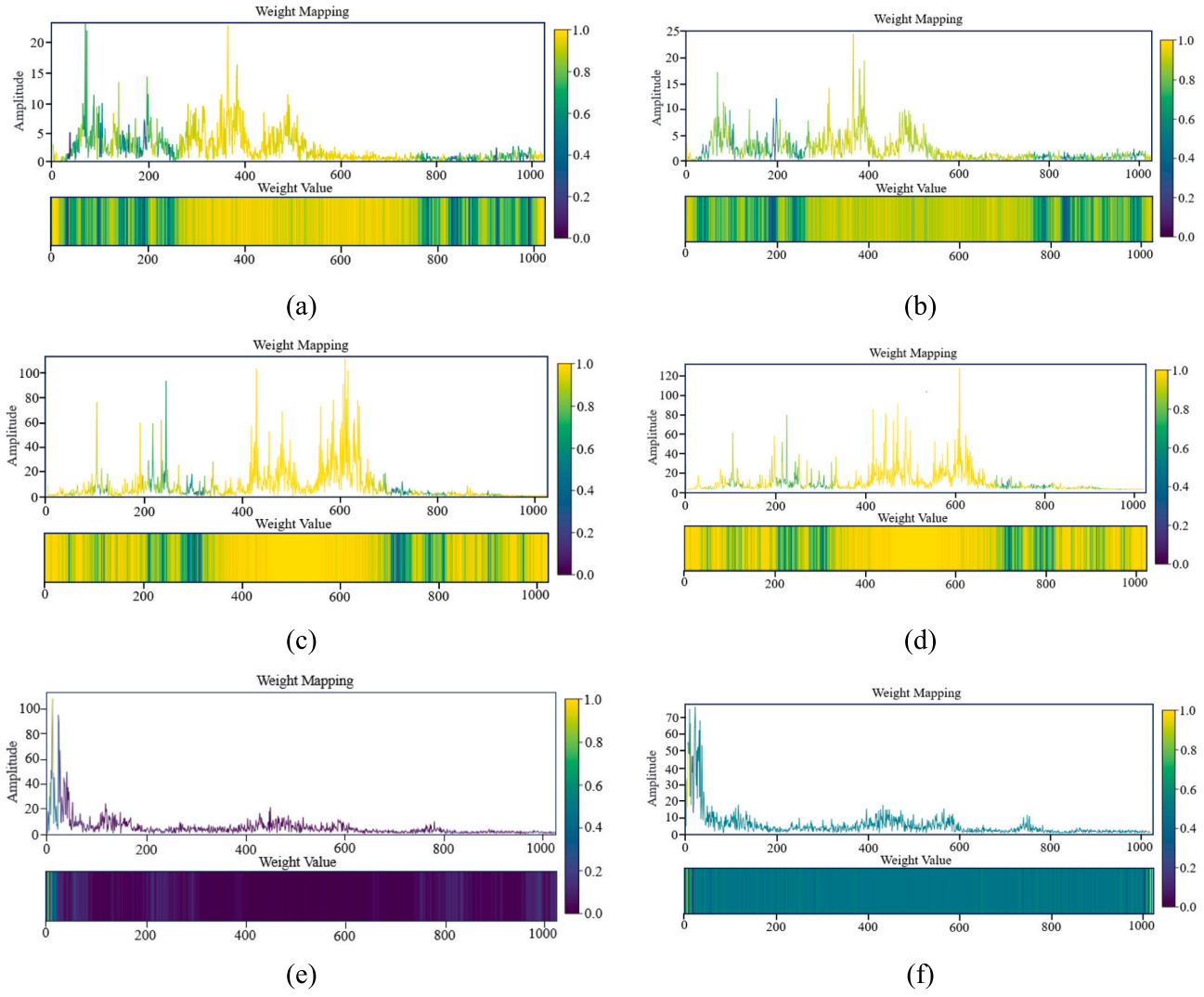


Fig. 14. Visualisation results of CAM. (a) CWRU-RF, (b) CWRU-OF, (c) LUT-RB, (d) LUT-OI, (e) RG-CT, (f) RG-HT.

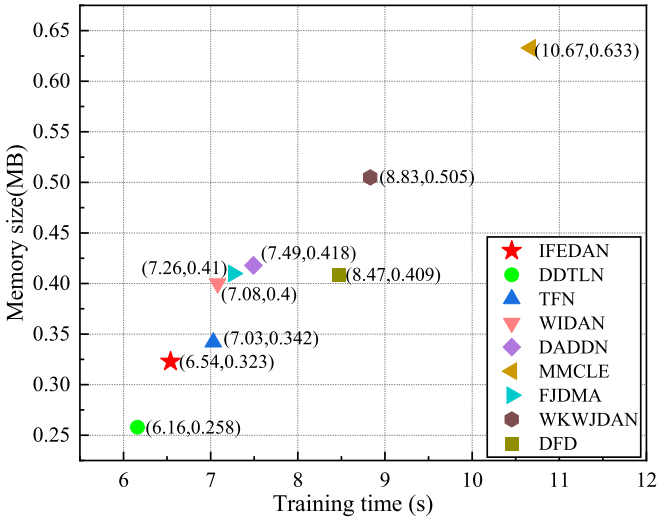


Fig. 15. The number of parameters and training time at 1 epoch.

4.6. Discussion

4.6.1. Timeliness and computing resources of IFEDAN

The computational resources and timeliness of IFEDAN are investigated, as shown in Fig. 15. In Fig. 15, DDTLN has the shortest training time, which is only 6.16 s. MMCLE has the longest training time, which is 10.67 s. In terms of the number of parameters, DDTLN has the fewest parameters, which is 0.258 MB, MMCLE has the highest number of parameters, which is 0.633 MB. This is because DDTLN has a simple structure, so the number of parameters is the lowest among all methods. In addition, WKWJDAN, WIDAN, FJDMA, and DFD all exhibit high the number of model parameters. However, the training time for the proposed IFEDAN is 6.54 s, and the number of parameters is 0.323 MB. Compared to other methods, IFEDAN has a shorter training time and a moderate number of parameters, demonstrating good training efficiency.

4.6.2. Parameter sensitivity analysis

There are adjustable parameters for IFEDAN, which affect the model's fault diagnosis performance. Therefore, parameter sensitivity analysis is conducted in this section, as shown in Fig. 16. Fig. 16 (a) shows the fault diagnosis results for different numbers of FERB, the horizontal axis represents the LUT and CWRU cross-domain diagnosis tasks, while the vertical axis represents the fault diagnosis accuracy rates

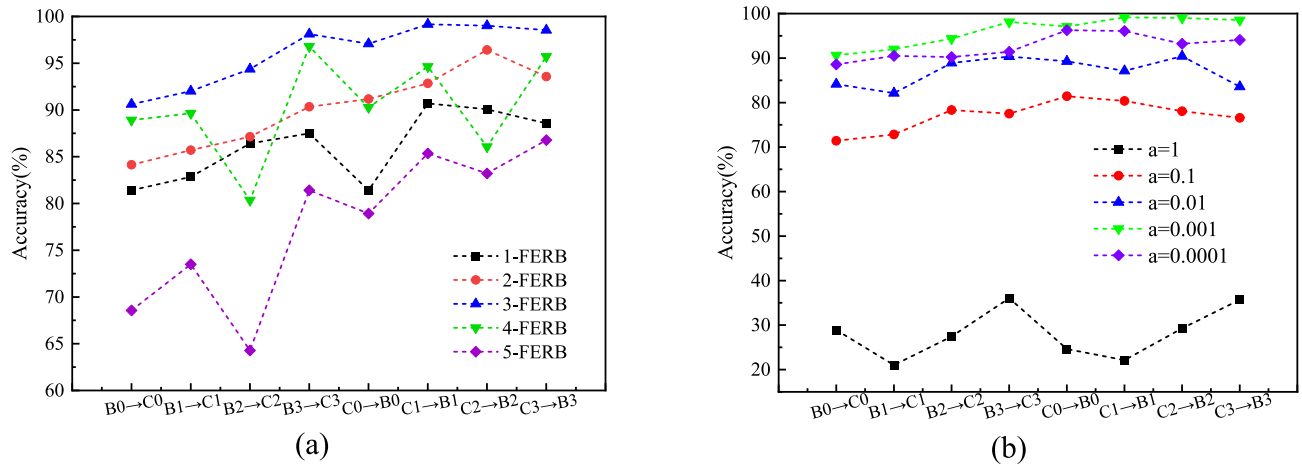


Fig. 16. Parameter sensitivity analysis. (a) Diagnostic results for different numbers of FERB, (b) Diagnostic results for different learning rates.

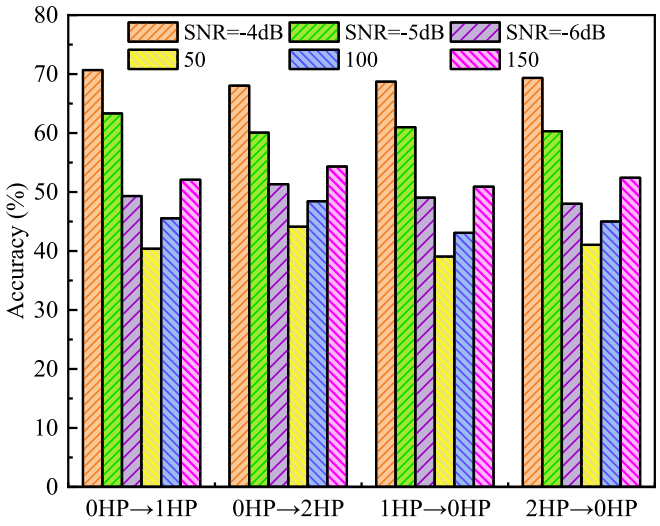


Fig. 17. Fault diagnosis results under strong noise environments and small sample conditions.

for different numbers of FERB. When FERB = 3, the fault diagnosis accuracy of IFEDAN is the highest, when FERB ≤ 3, the model's feature learning capability is insufficient, leading to poor fault diagnosis performance, when FERB ≥ 3, the performance of IFEDAN decreases sharply. This is because the excessive number of model parameters causes overfitting.

Fig. 16(b) shows the fault diagnosis results of IFEDAN under different learning rates. As can be seen from Fig. 16(b), when a = 1, the performance of IFEDAN decreases sharply. This is because the learning rate is set too high, causing the model to underfit during training. When a ≤ 0.1, the performance of IFEDAN is significantly improved. When a = 0.001, IFEDAN achieves the highest diagnostic accuracy.

4.6.3. Model limitations and failure scenarios

To guide future research, the limitations and failure scenarios of the model are investigated. The model is validated under stronger noise environments and small sample conditions. SNR values of −4 dB, −5 dB, and −6 dB are added. The small samples are 50, 100, and 150, and the training and test sets are divided according to 0.2:0.8. Fig. 17 shows the experimental results. As shown in Fig. 17, the model's fault diagnosis accuracy significantly decreases under strong noise environments and small sample conditions. This indicates the noise interference under strong noise environments blurs the fault features, limiting IFEDAN's feature extraction capability. Additionally, under small sample conditions, the limited number of training samples results in insufficient feature learning capability for IFEDAN. Therefore, in future work, we will integrate noise reduction techniques into IFEDAN to enhance the model's diagnostic performance under strong noise environments and small sample conditions.

5. Conclusion

In order to improve the cross-domain diagnostic performance of transfer learning methods under noisy environments and variable loads, we propose an interpretable frequency-enhanced domain adaptive network (IFEDAN) for cross-domain fault diagnosis of rotating machinery. Specifically, we construct the frequency enhanced residual feature extractor. This extractor introduces Morlet wavelet for weight initialization in the initial layer to enhance interpretability. Then, frequency enhanced residual block is constructed to help the model capture more transferable features and enhance useful features. In addition, Entropy Maximum Mean Difference (EMMD) loss is designed to enhance the stability and decision boundaries. Finally, validation is performed on a public dataset (CWRU) and self-built roller gear (RG) dataset and Lanzhou University of Technology (LUT) dataset. The results show that IFEDAN has good robustness and generalization performance. When diagnosing across domains between CWRU and LUT, the average diagnosis accuracy of IFEDAN is 96.11 %, which is higher than the comparison methods.

CRediT authorship contribution statement

Yazhou Zhang: Writing – original draft, Validation, Methodology. **Xiaoqiang Zhao:** Writing – review & editing, Validation, Methodology, Funding acquisition. **Zhenrui Peng:** Validation, Methodology, Funding acquisition. **Yongyong Hui:** Validation, Methodology. **Rongrong Xu:** Validation, Methodology. **Peng Chen:** Funding acquisition, Data curation.

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.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (No. 62263021), the College Industrial Support

Project of Gansu Province (2023CYZC-24), the Science and Technology Project of Gansu Province (24JRRA172), the Longyuan Young Innovative Talent Team Project of Gansu Province (310100296012), the Gansu Provincial Basic Research Innovation Group of China (25JRRA058), the Central Government's Funds for Guiding Local Science and Technology Development of China (25ZYJA040)

Data availability

Data will be made available on request.

References

[1] Liu X, et al. Novel source domain filtering dual classifier network with adaptive pseudo label refinement for partial domain fault diagnosis. Neurocomputing 2024:129260.
[2] Wang S, Zhao F, Cheng C, Chen H, Jiang Y. Threshold-optimized and features-fused semi-supervised domain adaptation method for rotating machinery fault diagnosis. Neurocomputing 2025;613:128734.
[3] Jiang Y, et al. Recursive prototypical network with coordinate attention: a model for few-shot cross-condition bearing fault diagnosis. Appl Acoust 2025;231:110442.
[4] Zhang Z, He H, Xu S, Yin L, Dong X. A reusable decoder network penalized by smooth group lasso and its applications to large-scale fault diagnosis of machinery. Control Eng Pract 2024;153:106127.
[5] Chen X, Yang R, Xue Y, Song B, Wang Z. TFPred: Learning discriminative representations from unlabeled data for few-label rotating machinery fault diagnosis. Control Eng Pract 2024;146:105900.
[6] Xiao Y, Shao H, Yan S, Wang J, Peng Y, Liu B. Domain generalization for rotating machinery fault diagnosis: a survey. Adv Eng Inf 2025;64:103063.
[7] Zhou X, et al. A hybrid denoising model using deep learning and sparse representation with application in bearing weak fault diagnosis. Measurement 2022;189:110633.
[8] Liu R, Ding X, Zhang Y, Zhang M, Shao Y. Variable-scale evolutionary adaptive mode denoising in the application of gearbox early fault diagnosis. Mech Syst Sig Process 2023;185:109773.
[9] Yang X, Yuan X, Ye T, Zhu W, Zhou F, Jin J. PSNN-TADA: prototype and stochastic neural network based twice adversarial domain adaptation for fault diagnosis under varying working conditions. IEEE Trans Instrum Meas 2023.
[10] Zhao X, Zhang Y. An intelligent diagnosis method of rolling bearing based on multi-scale residual shrinkage convolutional neural network. Meas Sci Technol 2022;33(8):085103.
[11] Liang H, Cao J, Zhao X. Average descent rate singular value decomposition and two-dimensional residual neural network for fault diagnosis of rotating machinery. IEEE Trans Instrum Meas 2022;71:1–16.
[12] Lv J, Xiao Q, Zhai X, Shi W. A high-performance rolling bearing fault diagnosis method based on adaptive feature mode decomposition and transformer. Appl Acoust 2024;224:110156.
[13] Zhang Y, Zhao X, Liang H, Chen P. Multiscale dilated convolution and swin-transformer for small sample gearbox fault diagnosis. Appl Intell 2024:1–17.
[14] Li S, et al. Dconformer: a denoising convolutional transformer with joint learning strategy for intelligent diagnosis of bearing faults. Mech Syst Sig Process 2024;210:111142.
[15] Zhang Y, Zhao X, Peng Z, Xu R, Chen P. WD-KANTF: an interpretable intelligent fault diagnosis framework for rotating machinery under noise environments and small sample conditions. Adv Eng Inf 2025;66:103452.
[16] Zhao B, Wu Q, Zhao K, Li J, Zhang Z, Shao H. A novel cross-receptive field fusion cascade network with adaptive mask update for transfer health state diagnosis of manipulators. Mech Syst Sig Process 2025;224:111976.
[17] He C, Shi H, Li J. IDSN: a one-stage interpretable and differentiable STFT domain adaptation network for traction motor of high-speed trains cross-machine diagnosis. Mech Syst Sig Process 2023;205:110846.
[18] Wu Z, Jiang H, Zhu H, Wang X. A knowledge dynamic matching unit-guided multi-source domain adaptation network with attention mechanism for rolling bearing fault diagnosis. Mech Syst Sig Process 2023;189:110098.
[19] Liu B, Yan C, He C, Lv M, Wei J, Wu L. An interpretable physics-informed subdomain moment-enhanced adaptation network for unsupervised transfer fault diagnosis of rolling bearing. Adv Eng Inf 2025;67:103491.
[20] Xiao Y, Shao H, Wang J, Cai B, Liu B. Domain-augmented meta ensemble learning for mechanical fault diagnosis from heterogeneous source domains to unseen target domains. Expert Syst Appl 2025;259:125345.
[21] Shao X, Cai B, Zou Z, Shao H, Yang C, Liu Y. Artificial intelligence enhanced fault prediction with industrial incomplete information. Mech Syst Sig Process 2025;224:112063.
[22] Wang R, Yan F, Yu L, Shen C, Hu X, Chen J. A federated transfer learning method with low-quality knowledge filtering and dynamic model aggregation for rolling bearing fault diagnosis. Mech Syst Sig Process 2023;198:110413.
[23] Pu Y, Tang J, Li X, Wei C, Huang W, Ding X. Single-domain incremental generation network for machinery intelligent fault diagnosis under unknown working speeds. Adv Eng Inf 2024;60:102400.
[24] Han T, et al. Novel adaptive loss weighted transfer network for partial domain fault diagnosis. ISA Trans 2024;145:362–72.
[25] Yan J, Cheng Y, Wang Q, Liu L, Zhang W, Jin B. Transformer and graph convolution-based unsupervised detection of machine anomalous sound under domain shifts. IEEE Trans Emerg Top Comput Intell 2024.
[26] Jiang X, Li X, Wang Q, Song Q, Liu J, Zhu Z. Multi-sensor data fusion-enabled semi-supervised optimal temperature-guided PCL framework for machinery fault diagnosis. Inf Fusion 2024;101:102005.
[27] Song Q, Jiang X, Liu J, Shi J, Zhu Z. Contrast-assisted domain-specificity-removal network for semi-supervised generalization fault diagnosis. IEEE Trans Neural Networks Learn Syst 2024.
[28] Yan J, et al. Multi-modal imitation learning for arc detection in complex railway environments. IEEE Trans Instrum Meas 2025.
[29] Wu H, Li J, Zhang Q, Tao J, Meng Z. Intelligent fault diagnosis of rolling bearings under varying operating conditions based on domain-adversarial neural network and attention mechanism. ISA Trans 2022;130:477–89.
[30] Yu X, Wang Y, Liang Z, Shao H, Yu K, Yu W. An adaptive domain adaptation method for rolling bearings' fault diagnosis fusing deep convolution and self-attention networks. IEEE Trans Instrum Meas 2023;72:1–14.
[31] Jiang L, Lei W, Wang S, Guo S, Li Y. A Deep convolution multi-adversarial adaptation network with correlation alignment for fault diagnosis of rotating machinery under different working conditions. Eng Appl Artif Intel 2023;126:107179.
[32] Yao Y, Chen Q, Gui G, Yang S, Zhang S. A hierarchical adversarial multi-target domain adaptation for gear fault diagnosis under variable working condition based on raw acoustic signal. Eng Appl Artif Intel 2023;123:106449.
[33] Shao H, Li W, Cai B, Wan J, Xiao Y, Yan S. Dual-threshold attention-guided GAN and limited infrared thermal images for rotating machinery fault diagnosis under speed fluctuation. IEEE Trans Ind Inf 2023;19(9):9933–42.
[34] Cheng Y, et al. Surrogate modeling of pantograph-catenary system interactions. Mech Syst Sig Process 2025;224:112134.
[35] Han Y, Lv S, Huang Q, Zhang Y. AMCW-DFFNSA: an interpretable deep feature fusion network for noise-robust machinery fault diagnosis. Knowl-Based Syst 2024;301:112361.
[36] Cheng Y, Zhou N, Wang Z, Chen B, Zhang W. CFFsBD: a candidate fault frequencies-based blind deconvolution for rolling element bearings fault feature enhancement. IEEE Trans Instrum Meas 2023;72:1–12.
[37] Ajakan H, Germain P, Larochelle H, Laviolette F, Marchand M. Domain-adversarial neural networks; 2014. arXiv preprint arXiv:1412.4446.
[38] Gao H, Zhang X, Gao X, Li F, Han H. Multi-timescale attention residual shrinkage network with adaptive global-local denoising for rolling-bearing fault diagnosis. Knowl-Based Syst 2024;304:112478.
[39] He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer; 2016. p. 630–45.
[40] Xu H, Zhou S. Maximum L-Kurtosis deconvolution and frequency-domain filtering algorithm for bearing fault diagnosis. Mech Syst Sig Process 2025;223:111916.
[41] Liang H, Cao J, Zhao X. Multibranch and multiscale dynamic convolutional network for small sample fault diagnosis of rotating machinery. IEEE Sens J 2023;23(8):8973–88.
[42] Wang Z, et al. Multi-modal multi-scale multi-level fusion quadrant entropy for mechanical fault diagnosis. Expert Syst Appl 2025;281:127715.
[43] Pang B, Liu Q, Sun Z, Xu Z, Hao Z. Time-frequency supervised contrastive learning via pseudo-labeling: an unsupervised domain adaptation network for rolling bearing fault diagnosis under time-varying speeds. Adv Eng Inf 2024;59:102304.
[44] Liu X, Liu F, Geng X, Fan L, Jiang M, Zhang F. Frequency domain guided latent diffusion model for domain generalization in cross-machine fault diagnosis. Measurement 2025:116989.
[45] Ragab M, et al. Conditional contrastive domain generalization for fault diagnosis. IEEE Trans Instrum Meas 2022;71:1–12.
[46] Qian Q, Qin Y, Luo J, Wang Y, Wu F. Deep discriminative transfer learning network for cross-machine fault diagnosis. Mech Syst Sig Process 2023;186:109884.
[47] Chen Q, et al. TFN: an interpretable neural network with time-frequency transform embedded for intelligent fault diagnosis. Mech Syst Sig Process 2024;207:110952.
[48] He C, Shi H, Liu X, Li J. Interpretable physics-informed domain adaptation paradigm for cross-machine transfer diagnosis. Knowl-Based Syst 2024;288:111499.
[49] Liu F, Deng W, Duan C, Qin Y, Luo J, Pu H. Duplex adversarial domain discriminative network for cross-domain partial transfer fault diagnosis. Knowl-Based Syst 2023;279:110960.
[50] Jin Y, Song X, Yang Y, Hei X, Feng N, Yang X. An improved multi-channel and multi-scale domain adversarial neural network for fault diagnosis of the rolling bearing. Control Eng Pract 2025;154:106120.
[51] Zhang Y, Zhao X, Xu R. Feature and joint distribution migration alignment method for cross-domain fault diagnosis of rotating machinery. IEEE Trans Instrum Meas 2025;74:3525115.
[52] Li X, et al. Multi-kernel weighted joint domain adaptation network for cross-condition fault diagnosis of rolling bearings. Reliab Eng Syst Saf 2025;261:111109.
[53] Liu Z, Zheng H, Liu H, Duan G, Tan J. A novel domain feature disentanglement method for multi-target cross-domain mechanical fault diagnosis. ISA Trans 2025.

Applied Intelligence (2024) 54:7716–7732
https://doi.org/10.1007/s10489-024-05530-x

Multiscale dilated convolution and swin-transformer for small sample gearbox fault diagnosis

Yazhou Zhang[1] · Xiaoqiang Zhao[1] · Haopeng Liang[2] · Peng Chen[3]

Accepted: 17 May 2024 / Published online: 13 June 2024

Abstract
Mechanical equipment usually operates in noisy and variable load environments, which presents serious challenges for existing intelligent diagnostic models. In addition, there are few labelled fault samples in real engineering scenarios, which makes it difficult to perform accurate fault identification for mechanical equipment. Thus, to solve the problem of diagnostic model performance degradation under small sample, noisy and variable load environments, this paper proposes a Multiscale Dilated Convolution and Swin-Transformer (MSDC-Swin-T) method for small sample gearbox fault diagnosis. First, we design the Coordinate Reconstruction Attention Mechanism (CRAM), which enhances the capture of impulse information by coordinate reconstruction. In addition, a multiscale convolutional token embedding module is constructed to extract local features at different scales, and its ability for capturing important features is adaptively enhanced by CRAM. Then, Swin-Transformer is utilized for modeling global dependencies, thus mining more subtle fault features. Finally, the effectiveness and stability of the MSDC-Swin-T is proved on two gearbox datasets. The experiments show that MSDC-Swin-T has superior diagnostic performance under small sample with noise and variable load environments. The diagnostic accuracy is better than the state-of-the-art methods.

Keywords Fault diagnosis · Gearbox · Dilated convolution · Swin-Transformer · Small sample

1 Introduction

With the rapid development of the Internet of Things, as the core components of various mechanical equipment, the mechanical transmission systems have started to develop in the direction of intelligence and automation [1–3]. Gearboxes, as the key components of the transmission system, tend to harbor a variety of failures due to the high intensity of work tasks. This would lead to unnecessary routine maintenance, property loss, and even serious safety accidents [4–6]. Therefore, intelligent fault diagnosis research for gearboxes is essential to ensure productivity and equipment reliability[7].

Conventional gearbox fault diagnosis methods usually rely on the priori experiences and signal processing techniques. Since these methods require specialized knowledge and ignore minor faults during fault detection, they lead to the difficulties in meeting the requirements of accurate fault identification and large-scale data processing [8, 9]. In recent years, many scholars have begun to utilize vibration data for fault identification, which is especially popular for intelligent fault diagnosis based on deep learning [10–13]. Deep learning methods can not only efficiently process a large amount of data and save a lot of manpower; but also automatically learn useful fault features without relying on expert knowledge. For example, Xiang et al. [14] proposed a fine topology and spatio-temporal GCN based fault diagnosis method for gearboxes. Ma et al. [15] developed a meta-learning framework based on multiscale dilated convolutional and relational modules. These methods use vibration data directly as model inputs, do not require expert knowledge, and achieve good diagnostic results. However, the quality and quantity of data are essential for deep learning methods. When the data samples are sparse or imbalanced in distribution, the performance would be degraded for deep learning-based fault diagnosis methods. In fact, it is very difficult to

✉ Xiaoqiang Zhao
xqzhao@lut.edu.cn

[1] College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China

[2] College of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China

[3] College of Electrical and Electronic Engineering, Lanzhou Petrochemical University of Technology, Lanzhou 730050, China

collect numerous fault samples in real engineering scenarios. For example, since the gearboxes of wind turbines usually do not fail within a short period of time, it may take months or even years for researchers to obtain fault data. Hence, it is still a challenging problem to establish an intelligent gearbox diagnostic model under small sample.

Currently, the mainstream methods for dealing with small sample in fault diagnosis are divided into three types: data augmented methods, transfer learning methods and few-shot learning methods. Data augmented methods increase the training data using techniques such as flipping, panning and cropping. For example, Zhang et al. [16] first used resampling technique to achieve sample expansion and then proposed an adaptive convolutional network to achieve fault diagnosis under small sample. Zhang et al. [17] proposed a generative adversarial network for small sample fault diagnosis. The method utilized data generation to solve the problem with insufficient number of samples. Transfer learning refers to the use of already learned features to improve the problem of insufficient feature extraction in new tasks. For example, Liu et al. [18] developed an agnostic meta-learning cross-domain diagnosis method, which ultimately achieved small sample fault diagnosis by constructing a feature extractor using multicore efficient channel attention. Ma et al. [19] proposed an adaptive unsupervised small sample diagnostic network with enhanced transferability, which designed a transferability strategy to enhance feature learning on the target domain. The methods of few-shot learning mean that the model accomplishes feature learning with a limited number of training samples to achieve good performance in new class or task. It aims to learn effectively in the data sparse situation. For example, Zheng et al. [20] proposed a few-shot diagnostic network based on improved meta-relational networks which designed a multiscale feature encoder to extract the features in the support and query, and then obtained the query samples by scoring between the support and query, and ultimately achieved fault classification. Shi et al. [21] proposed a transferable siamese network to obtain the distance between different samples, and to evaluate the similarities and differences between the samples. Although all these methods mentioned above have achieved encouraging results in fault diagnosis with small samples, there are still some issues that hinder their application. For example, the use of overlapping sampling results in redundancy of sample information. In addition, when the quality of the original data is poor, the performance of data augmentation methods based small sample fault diagnosis will be decreased. The performance of transfer learning relies on the model's ability to learn samples from the source domain. When the similarity is low between the source domain and the target domain, the fault diagnosis accuracy is poor. The network structure of few-shot learning is complex and are susceptible to the number of samples, resulting in slow training and poor generalization performance. In addition, the above study does not consider the strong noise interference in industrial production. Under strong noise, the original data would become extremely unstable, and some periodic features would be submerged, which lead to the difficulties for extracting fault information. Therefore, there is a need to develop an efficient and easily applied fault diagnosis method under small sample, which can not only accurately extract fault information, but also has stronger resistance to noise interference.

In recent years, Transformer has received great attention from scholars in the field of fault diagnosis[22–24]. For example, Ding et al. [25] proposed a new time–frequency Transformer to achieve fault diagnosis of bearings. Wu et al. [26] used the Transformer as a classifier, which not only achieved fault detection, but also efficiently identified different types of faulty conditions with different levels of severity. In addition, since the Transformer's self-attention mechanism can effectively capture the long-range dependence of global contextual information, it has been applied to noise immunity research. For example, Han et al. [27] proposed a convolutional transformer to extract global and local information to enhance fault diagnosis performance under strong noise. Li et al. [28] have developed a denoising convolutional transformer. It can extract local and global discriminative features from harsh environments, thus improving the network's anti-interference performance.

The above studies show that although the transformer has been employed successfully in the field of fault diagnosis, the performance of transformer heavily depends on the quality of the labelled samples. If the quantity of labelled samples is small and there is heterogeneity between the data, it would affect the training effect of the transformer and thus reduce the accuracy of fault diagnosis. In addition, using the transformer only as an encoder for feature extraction would lose the low-order local features of the original data. Especially, these low-order local features have a huge impact on the diagnostic performance of the model under small sample. Thus, this paper proposes a new gearbox fault diagnosis method based on multiscale dilated convolution and Swin-Transformer (MSDC-Swin-T) to make the transformer model have better low-order local feature extraction capability. The proposed method is different from traditional fault diagnosis methods. It is an end-to-end intelligent diagnostic model that does not require signal processing techniques. Firstly, a multiscale convolutional token embedding module is designed to extract the low-order local features at different scales from the input samples. This module uses dilated convolution for multiscale feature extraction and enhances the ability to capture critical impulse features by coordinate reconstruction attention mechanism. Then, the extracted low-order local features are segmented into fixed-size tokens and fed into the Swin-Transformer for modeling global dependencies. Finally, we validate the proposed method by different gearbox datasets, and the experimental results indicate that MSDC-Swin-T has significant advantages and high diagnostic accuracy

compared to existing state-of-the-art methods, especially, the problem of fault identification for mixed faults and small sample is well solved. The main contributions of this paper are summarized as follows:

(1) We propose a fault diagnosis method of gearboxes based on multiscale dilated convolution and Swin-Transformer. This method can solve the problem of degraded diagnostic performance for gearboxes under small sample, noisy and variable load environments.
(2) We design a coordinate reconstruction attention mechanism and use it as a basic component to construct a multiscale convolutional token embedding module. This module can help the Swin-Transformer to perform low-order local feature extraction, making it more adaptable to small sample fault diagnosis tasks. In addition, the Swin-Transformer module is subjected to token sparsity operations to overcome the overfitting phenomenon in the late stage of training the model.
(3) We conducted two case studies using the Southeast University gearbox dataset (spur gears) and the roller gear dataset (bevel gears). The results show that MSDC-Swin-T has good robustness under small sample and noisy environments, and has good generalization under small sample and variable load environments.

The rest of the paper is organized as follows. Section 2 briefly introduces the theories related to dilated convolutions and transformers. Section 3 introduces the proposed method. Section 4 carries out the experimental validation and analyze the results about the proposed method. Section 5 summarizes the relevant conclusion and looks forward to future research directions.

2 Relevant theoretical background

This section introduces the basic theory of dilated convolution. In addition, transformers that are applied to the fields of natural language processing (NLP) and computer vision are also introduced.

2.1 Dilated convolution

In CNNs, the standard convolution has the same size of receptive field as its filter size. Therefore, down-sampling operations are used to increase the network depth and decrease the number of parameters. However, dilated convolution uses the dilated rate to control the size of the receptive field, thus not only ensuring that the number of parameters of the network does not grow, but also that the receptive field increases with an exponential form [29, 30], and its receptive field is computed as:

$$F_{size} = k_s + (k_s - 1) \times d_{rate} \tag{1}$$

where k_s denotes the size of the dilated convolution, d_{rate} denotes the dilated rate, and F_{size} denotes the receptive field size. Figure 1 shows the schematic diagram of dilation convolution. In Fig. 1(b), when the dilated rate is 2, the receptive field of the dilated convolution increases significantly. However, the dimensions of its output features remain unchanged.

2.2 Transformer

Vaswani et al. [31] proposed Transformer in 2017. Nowadays it has become a benchmark for NPL tasks. The self-attention mechanism in Transformer not only effectively characterizes the global dependency between inputs and outputs, but also improves the overall learning capability by parallelizing the training process. Transformer has an encoder-decoder structure that consists of a position encoding, encoder and decoder as shown in Fig. 2. Among them, the encoder and decoder consist of multi-head self-attention (MSA), feed-forward network (FFN), layer normalization and residual connection.

2.2.1 Position encoding

Transformer uses self-attention to obtain relationships between input sequences. This results in sequence position information being ignored during parallel computation. Therefore, to compensate for this shortcoming, we perform

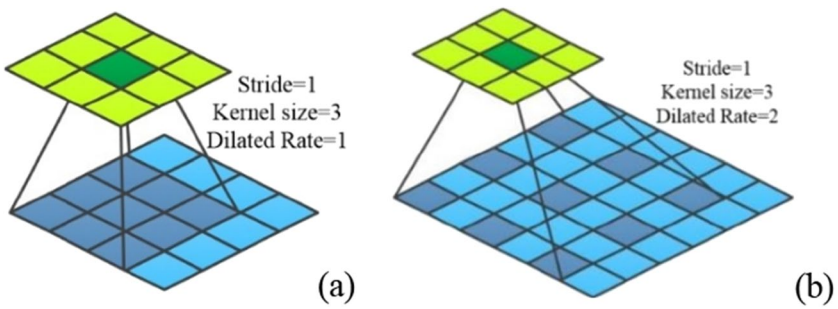


Fig. 1 Schematic diagram of dilated convolution. (a) dilated rate: 1, (b) dilated rate: 2

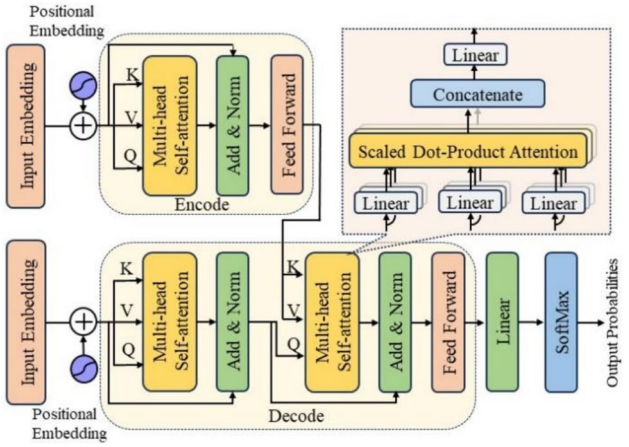


Fig. 2 Schematic of Transformer principle

positional encoding before self-attention. The position encoding is generally constructed by sine and cosine functions of different frequencies, which is described as:

$$EN_{(p,2i)} = sin(p \times \frac{1}{10000^{2i/d}}) \tag{2}$$

$$EN_{(p,2i+1)} = cos(p \times \frac{1}{10000^{2i/d}}) \tag{3}$$

$$\widehat{X} = X + EN \tag{4}$$

where $X = \{x_1^d, x_2^d, ..., x_n^d\}$, d denote the dimension of each vector of the set X. p denotes the input sequence position number, i is the dimensioned labeling position, and $\widehat{X}$ is the output after position encoding.

2.2.2 Multihead self-attention (MSA)

MSA is an important part of Transformer, which obtains important information in the input sequence by placing different weights at each position. In MSA, first, a linear mapping is performed on the token T of dimension d to generate the key matrix K, the query matrix Q and the value matrix V, respectively. Then the matrices K, Q and V are mapped to the projection set Q_i, K_i, $V_i \in \mathbb{R}^{d,d_h}$, $i = \{1,2,...,h\}$ by using the three sets of parameter matrices w_i^q, w_i^k and w_i^v, and where $d_h \times h = d$. Thus, the attention function of the i-th head can be defined as:

$$Att(T,i) = softmax(\frac{Q_i K_i^T}{\sqrt{d_h}})V_i \tag{5}$$

$$Q_i = w_i^q Q,\ K_i = w_i^k K,\ V_i = w_i^v V \tag{6}$$

The results of the attention computation are fused and linearly mapped to obtain the final output, which is described as:

$$MSA(T) = Linear\{concat_{i=1}^{h}(Att(T,i))\} \tag{7}$$

where $Linear\{\cdot\}$ denotes the linear mapping and $concat(\cdot)$ is the series operation.

2.2.3 Feed-forward layer

Each encoder and decoder of Transformer contains a feed-forward layer, which is obtained by doing two linear transformations of the rectified linear unit (ReLU). The process is described as:

$$F(X) = f_{relu}(0, Xw_1 + b_1)w_2 + b_2 \tag{8}$$

where w_1, w_2, b_1, b_2 denote the weight matrix and bias of the linear projection of the two layers, respectively. $f_{relu}(0, Xw_1 + b_1)$ denotes the linear projection of the first layer using Relu function.

2.2.4 Transformer encoder

Transformer encoder usually consists of multihead self-attention and feed-forward layers. In addition, the output of the encoder is optimized by residual connection and layer normalization. In the transformer structure, global perceptual features are captured by stacking multiple encoders to meet complex task requirements. The process is described as:

$$X_M = F_{layer}^{norm}(MSA(X) + X) \tag{9}$$

$$X_{FFL} = F_{layer}^{norm}(FFL(X_M) + X_M) \tag{10}$$

where X denotes the input of the encoder, X_M denotes the output of the self-attention, and X_{FFL} denotes the output of the encoder.

3 Proposed MSDC-Swin-T

The proposed method consists of multiscale convolutional token embedding module, Swin-Transformer with Sparse Tokens and classifier. First, the samples are acquired through a time-shifted window and fed into the multiscale convolutional token embedding module to extract low-order local feature information. Then, to improve the models' fine-grained observation capability and modeling global dependencies, Swin-Transformer is used to construct mapping relationships between mixed faults and features under small sample with noise and variable load environments. Finally, classifier is used for fault diagnosis.

3.1 Multiscale convolutional token embedding module

3.1.1 Coordinate reconstruction attention mechanism

In practice, the vibration signals of gearboxes are usually non-stationary, and the strong noise and varying loads in the operating environment further increase the non-linear characteristic of the vibration signals, leading to the difficulty of effectively capturing the potential fault information. Therefore, it is critical to enhance the feature extraction capability of the model. Although the coordinate attention mechanism can encode positional information and help the model to better focus on the features at different locations, the mechanism is mainly applied in target detection and image segmentation [32]. In order to effectively apply the coordinate attention mechanism to vibration signal analysis, this paper designs a coordinate reconstruction attention mechanism that aims to provide position information encoding that is more applicable to vibration signal features. It is described in detail as shown in Fig. 3.

The coordinate reconstruction attention mechanism can be divided into the following steps.

Step1: The dimensionality of the input features is compressed using Global Maximum Pooling (GMP). This is because Global Maximum Pooling (GMP) can capture the critical impulse information. The process is described as:

$$X_c^{gmp} = \underset{0\le j<d}{Max}\{x_c(1,j)\} \tag{11}$$

where X_c^{gmp} is the c-th channel output feature, and $Max\{\cdot\}$ is the maximum pooling operation.

Step2: To capture the spatial location information, we use the convolution operation to fuse X_c and X_c^{gmp} together. The process is described as:

$$\widehat{X} = Cat(f_{conv}(X_c, X_c^{gmp})) \tag{12}$$

where $Cat(\cdot)$ is the fusion operation, and $f_{conv}(\cdot)$ is the convolution operation.

Step3: To make CRAM more applicable to small sample and increase its ability to capture critical impulse features, we design the adaptive sigmoid activation function. Specifically, first, we perform linear transformations by 1×1 convolutional layer, and utilize a normalization operation for normalizing the convolutional output, then two learnable weight matrices P_1 and P_2 are defined, and multiplied with the input features. The learnable weight matrix is optimized during model training, making the model more suitable for small sample tasks. Finally, sigmoid activation function is used to limit the feature values between 0 and 1. The adaptive sigmoid activation function optimizes the weights of the fused feature information. The process is described as:

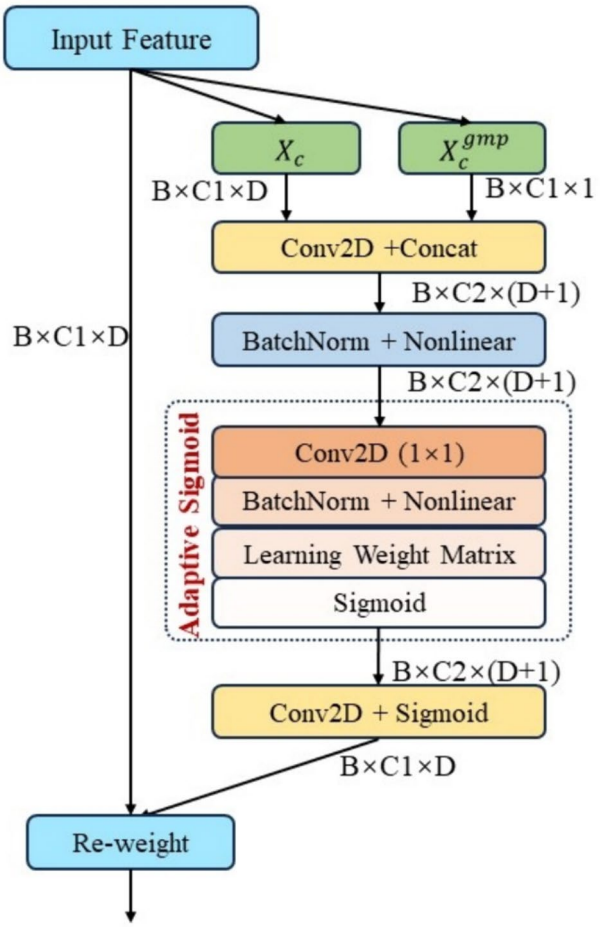


Fig. 3 Coordinate reconstruction attention mechanism

$$F = \delta_{Adp}(\widehat{X}) \tag{13}$$

where $\delta_{Adp}(\cdot)$ is the adaptive sigmoid activation function.

To further increase the expressiveness of F, we perform linear transformations by using the sigmoid function and the convolution operation. The process is described as:

$$\widehat{F} = \delta(f_{conv}(F)) \tag{14}$$

Step4: $\widehat{F}$ is residually connected to the input feature X_c. Thus, the output of CRAM is characterized not only with pulse information, but also with the original signal features. The process is described as:

$$Y = X_c \otimes \widehat{F} \tag{15}$$

where Y is the CRAM output feature and $\otimes$ is the residual connection.

In the coordinate reconstruction attention mechanism, the global maximum pooling is used to capture the impulse informations from the input. These impulse informations

contain periodic fault features of gears, which can improve the diagnostic performance of the model under small sample and noisy environments. In addition, we design an adaptive sigmoid activation function, which can relocate the input features and encode the features after global maximal pooling, thus reducing the effect of attentional bias. Compared with residual networks, the coordinate reconstruction attention mechanism can effectively encode the location of fault information so that it can better focus on fault information at different locations and improve the performance of the model. The residual network introduces residual connections to help the network learn the constant mapping more easily and accelerate the network convergence speed.

3.1.2 Multiscale convolutional token embedding module

A typical Swin-Transformer splits the input samples into multiple fixed-size patches and preserves the positional information of each patch. Then, each patch is mapped to a token by linear projection. However, the tokens via direct linear projection cannot capture low-level local features, resulting in degradation of the transformer's performance in harsh environments. In addition, the vibration signals from gearboxes exhibit non-stationary characteristic, and there are differences in the fault features contained in different time scales. Therefore, it is critical to ensure that the model can effectively extract multiscale low-order local features. To address the above shortcomings, we design a multiscale convolutional token embedding module which aims to extract sufficient low-order local features from vibration signals to improve the fault diagnosis performance of the model. Specifically, the multiscale convolutional token embedding module first performs multiscale feature extraction using dilated convolutions with different dilated rates. Then, the extracted features at different scales are positionally encoded using the coordinate reconstruction attention mechanism to focus on useful features and suppress redundant features. Finally, the extracted features at different scales are fused and used as inputs for the next module. The schematic diagram of the multiscale convolutional token embedding module is shown in Fig. 4.

In Fig. 4, first, we resample the vibration signals using the moving window of 784. Then, dilated convolution is conducted for multi-scale feature extraction, where the number of channels in the dilated convolution is 96 and dilated rates are 1, 2 and 5, respectively. In addition, we optimize the feature extraction process using batch normalization and the Relu function. Finally, the intrinsic correlation of features is learned using the coordinate reconstruction attention mechanism and the low-order local features extracted from different scale convolutions are fused, which are used as the inputs to the Swin-Transformer module.

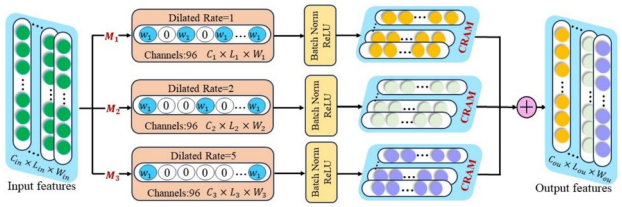


Fig. 4 Multiscale convolutional token embedding module

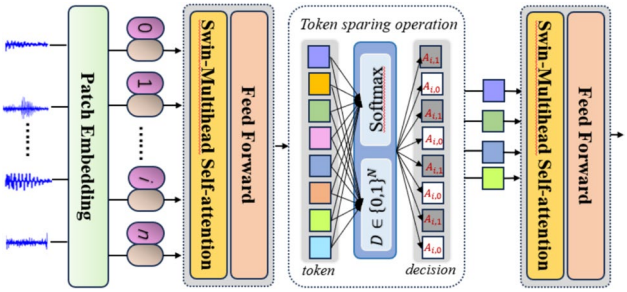


Fig. 5 Schematic diagram of token sparing operation

3.2 Swin-transformer with sparse tokens

Fault features of gearboxes exhibit periodic characteristic. In noisy and variable load environments, these periodic fault features are prone to be disturbed and difficult to be recognized effectively. The Swin-Transformer utilizes the multihead self-attention mechanism and positional encoding which can capture long-range dependencies, enabling it to better mine the periodic fault features of gearboxes. Therefore, after the multiscale convolutional token embedding module, we use Swin-Transformer for higher-level feature mining to improve the fault diagnosis performance of the model. However, from our experiments we have found that increasing the number of Swin-transformers leads to unstable training process, which reduces the classification accuracy. Specifically, although increasing the number of layers of Swin-Transformer improves feature extraction, it also increases redundant information and exacerbates the overfitting problem with small sample. Thus, we design the token sparse operation as shown in Fig. 5.

In Fig. 5, first we evaluate all the tokens in each layer, which retain the useful token information and discard the redundant token information. Specifically, we define the decision mask $D_n \in \{0,1\}$, where n is the number of all tokens. All elements of the decision mask have the initial value 1. Second, the decision mask D and all tokens $N = \{X \in \mathbb{R}^{H\times W\times C}\}$ are used as the inputs, and the local features of the tokens are computed. The process is described as follows:

$$Z_{local} = F(X) \in \mathbb{R}^{h\times w\times c'} \tag{16}$$

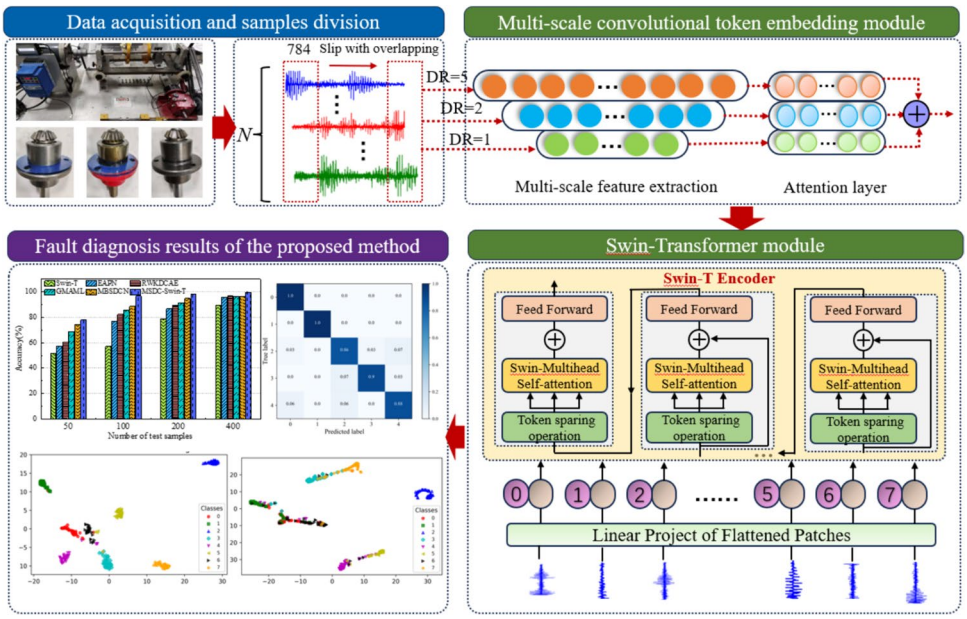


Fig. 6 Diagnostic flow of the proposed method

where h is the length of the token, w is the width of the token, and $c' = c/2$ is the half number of channels. In addition, the global features of the token are computed. The process is described as:

$$Z_{global} = Cat(F(X), D) \in \mathbb{R}^{h \times w \times c} \tag{17}$$

$$Cat(x, D) = \frac{\sum_{i=1}^{N} D_i x}{\sum_{i=1}^{N} D_i}, x \in X \tag{18}$$

Finally, we fuse the local and global features. The calculation of discarding or retaining the token is performed using Softmax function. The process is described as follows:

$$Z = [Z_{local}, Z_{global}] \tag{19}$$

$$A = Softmax(Z) \tag{20}$$

where, $A_{i,0}$ is the i-th discarded token and $A_{i,1}$ is the i-th retained token. We use A to generate the current decision mask and update its value. The process is described as:

$$\hat{D} \leftarrow (D \odot \overline{)D}) \tag{21}$$

where $\overline{)D}$ is the discarded token.

3.3 Rectified adam (Radam)

In the early stages of model training, the larger first-order moment estimates of the Adam optimizer can help the model converge quickly. However, the large first-order moment estimate causes the model to oscillate or unstable in the later stages of training. Radam optimizer uses a correction term in the computation of the gradient

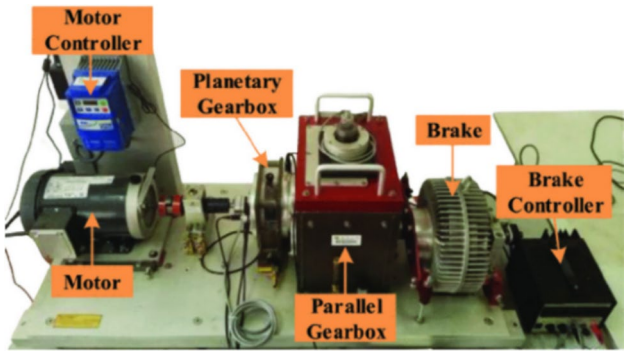


Fig. 7 Gearbox test rig

momentum. The adjustment term can readjust the gradient momentum based on current and past values [33]. The initial values of the gradient momentum and the second-order momentum of the gradient are defined as follows:

$$\begin{cases} m_0 = 0 \\ v_0 = 0 \end{cases} \tag{22}$$

Adam optimizer updates the values of the gradient momentum and the second-order momentum of the gradient at time where g_t is the value of the gradient at time t, and β_1 and β_2 are the exponential decay rates of the moment estimates, generally $\beta_1 = 0.9$, $\beta_2 = 0.999$.

Radam optimizer updates the values of the gradient momentum and the second-order momentum of the gradient at moment t.

$$\begin{cases} m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t \\ v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2 \end{cases} \tag{23}$$

Table 1 Description of dataset failure types

Location	Type	Fault description	Label
Bearing	Health_1	Normal condition	0
	Ball	Ball race fault	1
	Inner	Inner race fault	2
	Outer	Outer race fault	3
	Combination	Inner and outer race fault	4
Gearbox	Health_2	Normal condition	5
	Chipped	Chipped tooth in the gear	6
	Miss	Missing tooth in the gear	7
	Root	Gear root cracking	8
	Surface	Gear surface wear	9

$$\widehat{m}_t = \frac{m_t}{1-\beta_1^t} + \frac{(1-\beta_1^t)\cdot\gamma}{(1-\beta_1^t)\cdot\gamma+\beta_1^t} \tag{24}$$

$$\widehat{v}_t = \frac{v_t}{1-\beta_2^t} + \frac{(1-\beta_2^t)\cdot\gamma}{(1-\beta_2^t)\cdot\gamma+\beta_2^t} \tag{25}$$

where γ is the adjustable hyperparameter. From Eqs. (24) and (25), Radam optimizer adds a correction term in the process of updating the values, thus suppressing the problem of excessive growth of first-order moment estimates. Finally, the parameter update process is described as:

$$w_{t+1} = w_t - \alpha \bullet \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t}+\varepsilon} \tag{26}$$

where $\alpha = 0.0001$ is the learning rate and is the bias.

3.3.1 Gradient cropping

During model training, Radam constantly calculates the gradient of the model parameters and then continuously updates the parameters based on the gradient. However, the gradient sometimes becomes very large or very small, which leads to unstable training. Therefore, we use the gradient cropping technique to adjust the size of the gradient to ensure the stability of the gradient. Specifically, gradient cropping prevents the gradient to become too large during the training process by limiting the paradigms of the gradient. In this paper, we set the threshold to 0.5. When the paradigms of the gradient exceed this threshold, the gradient is shrunk so that it does not exceed the threshold, thus avoiding the gradient explosion problem.

3.4 MSDC-Swin-T diagnostic process

The general architecture of the proposed method is shown in Fig. 6, and is described as follows:

(1) The signals are collected from different gear health states under different operating conditions. In addition, a fixed-size moving window is used to divide the model input samples.
(2) In the multiscale convolutional token embedding module, low-order local features are extracted by dilated convolutions with different sizes of dilated rates, and the important features is enhanced by CRAM.
(3) Swin-Transformer with sparse tokens is used for deeper feature learning, and multi-head self-attention mechanism for fine-grained feature extraction.
(4) MSDC-Swin-T diagnostic model is trained using a limited number of labeled samples. Test samples are fed

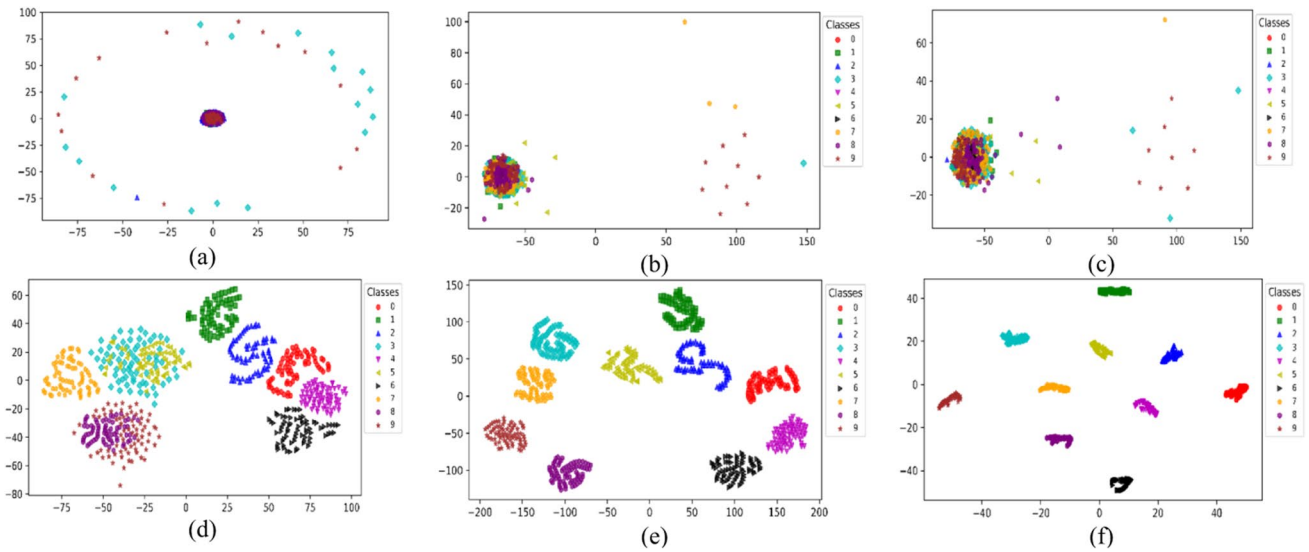


Fig. 8 t-SNE visualization of key layer output features. (a) Original layer (b) Dilated convolution layer (c) CRAM layer (d) Swin-T first layer (e) Swin-T second layer (f) Output layer

Table 2 Parameters of the Swin-Transformer with Sparse Tokens

Hyperparameter	Value	Hyperparameter	Value
Patch size	2	Dropout after MLP layer	0.01
Number of attention heads	8	Dropout after Attention	0.01
Number of embedded dimensions	64	Drop path rate	0.1
Number of MLP nodes	256	Number of Swin-T blocks	2

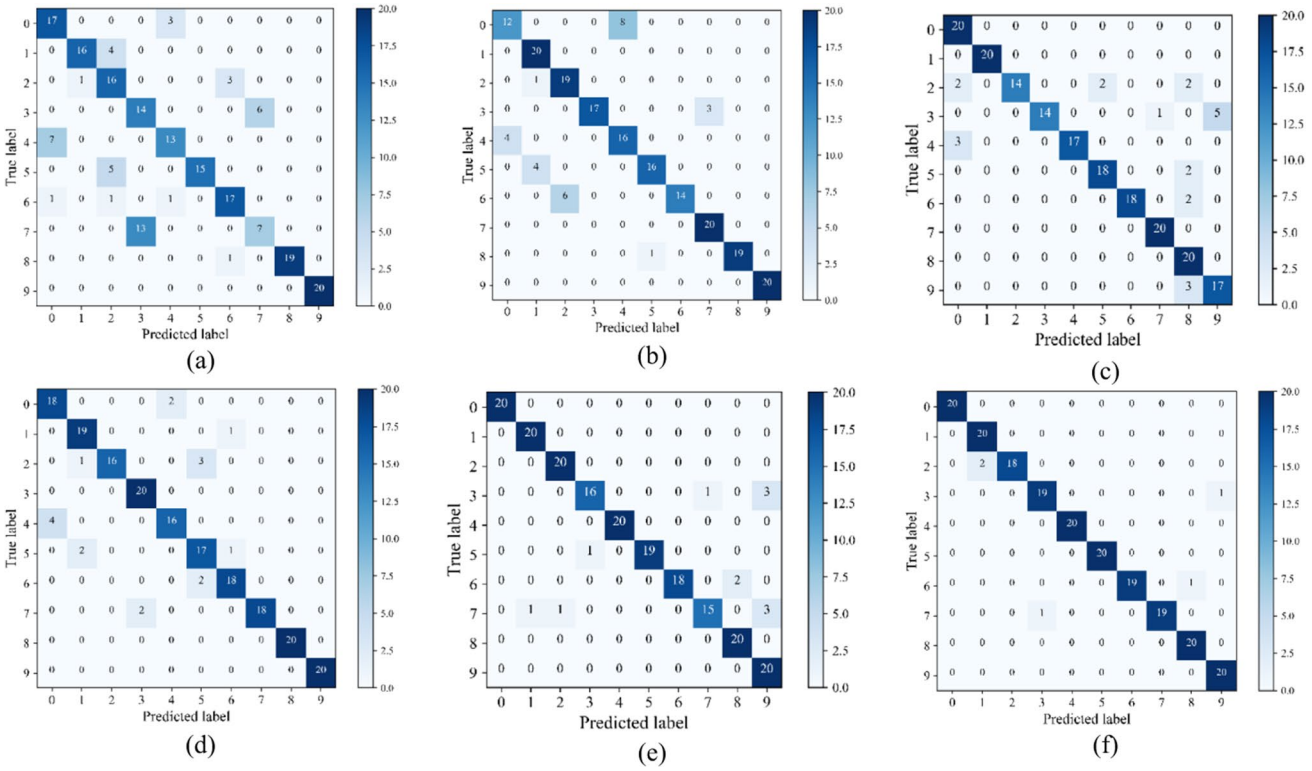


Fig. 9 Confusion matrix classification results. (a) Swin-T (b) EAPN (c) RWKDCAE (d) GMAML (e) MBSDCN (f) MSDC-Swin-T

into the completed trained MSDC-Swin-T to validate its performance.

4 Experimental verification and analysis

To validate the effectiveness of the proposed method under small sample, noisy and variable load environments, two case studies are conducted using the Southeastern University gearbox dataset (spur gears) and the roller gear dataset (bevel gears) in this section.

In two case studies, four state-of-the-art methods and a Swin-Transformer (Swin-T) benchmark model are used to assess the superiority of the proposed method: Effective attention prototype network (EAPN) [34], Residual wide kernel deep convolutional autoencoder (RWKDCAE) [35], Generalized model agnostic meta-learning (GMAML) [18], and Multi-branch multiscale dynamic convolutional network (MBSDCN) [36]. For the fair comparison, the comparison methods are aligned with the proposed method in terms of the number of network layers. Among them, EAPN consists of effective lightweight channel attention and dilated convolution. RWKDCAE consists of wide kernel convolutional layers, residual learning and convolutional autoencoder. GMAML consists of multi-kernel effective channel attention and meta-learning strategy. MBSDCN consists of multiscale convolution and channel reconstruction attention.

4.1 Case 1

4.1.1 Description of the dataset

In this case, the gearbox dataset from Southeast University in China is used, and the fault data acquisition test rig is shown in Fig. 7. The test rig consists of a motor, two gearboxes and motor controller. Type 608A11 vibration sensors were used to collect

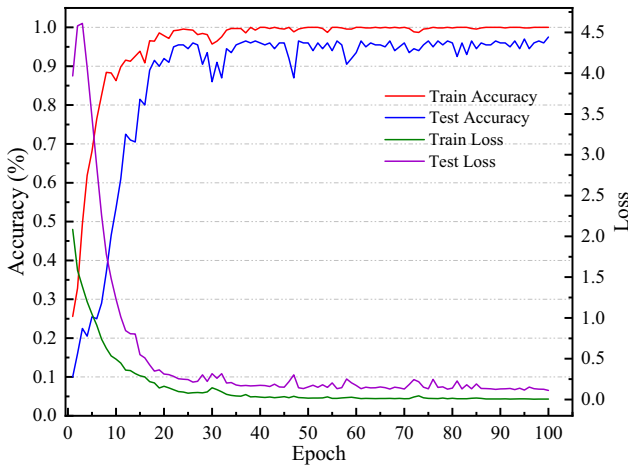


Fig. 10 Convergence performance of the proposed method

vibration data from gearboxes in the X, Y, and Z directions and on the Z-axis of the motor. The objects to be tested were spur gears and rolling bearings in this test rig. The bearing and gear data were obtained by adjusting the motor speed and load for both 20 Hz-0 V and 30 Hz-2 V operating conditions. Among them, the gear and bearing data contain five fault states respectively. We select the gear and bearing data under the 20 Hz-0 V condition to verify the classification performance of the proposed method. The description of the dataset is shown in Table 1.

The samples were divided using the sliding window, each type of fault samples is 100, and the total number of samples is 1000. The numbers of training samples, tests samples and validation samples are in the ratio of 7:2:1.

4.1.2 Parameters and visualization analysis of the proposed method

The parameters of the Swin-Transformer with Sparse Tokens are shown in Table 2, these parameters are mainly determined by cross-validation. In addition, to visualize the feature extraction capability of the proposed method, we use t-SNE to visualize the features extracted for each key layer, as shown in Fig. 8. In Fig. 8(a), most of the features are mixed together and only a few features are distributed at the edge positions. In Fig. 8(b) and (c), the mixed features start to move towards the edge positions after the multiscale convolutional embedding module, but this trend is not so obvious. The distance between different fault features is further expanded after CRAM layer processing, and the tendency of different features to spread towards the edge is strengthened. Then, the similar fault aggregation effect becomes more obvious with the depth of the network, as shown in Fig. 8(d) and (e). Particularly, Swin-Transformer layer represents the stronger aggregation effect than the previous layers, which means that the multi-head attention of this layer can pay more attention to the subtle features and improve the fault diagnosis accuracy of the proposed method. Finally, similar faults are fully aggregated in the output layer, and there is no overlap for different classes of faults, which indicates that the proposed method has a strong feature extraction capability.

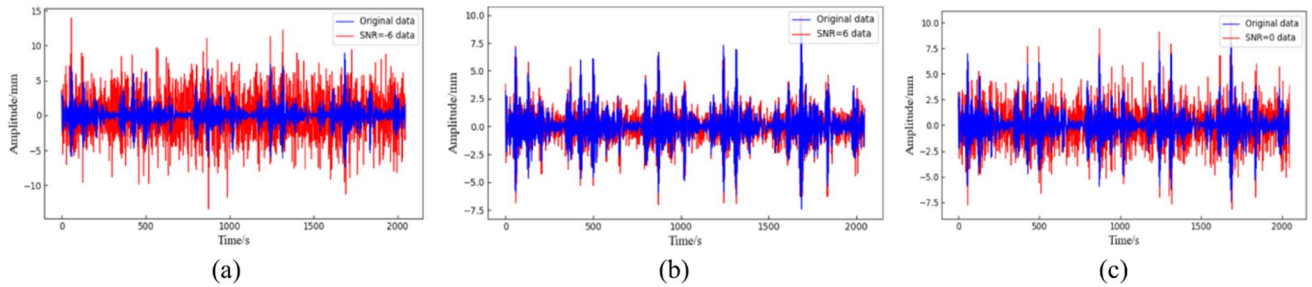


Fig. 11 Different SNR added to the original signal. (**a**) SNR = -6 dB (**b**) SNR = 0 dB (**c**) SNR = 6 dB

Table 3 Accuracy of the six methods in noise environments

Algorithms	Accuracy (%)			
	-4 dB	0 dB	4 dB	8 dB
Swin-T	55.92 ± 2.12	64.76 ± 2.67	71.50 ± 1.53	74.54 ± 1.78
EAPN	58.50 ± 1.87	74.00 ± 1.76	82.04 ± 0.23	84.34 ± 2.34
RWKDCAE	64.21 ± 1.01	78.92 ± 2.46	83.67 ± 0.87	86.76 ± 0.74
GMAML	78.00 ± 0.17	77.50 ± 1.83	88.37 ± 1.03	90.12 ± 0.32
MBSDCN	81.24 ± 0.57	88.00 ± 0.31	92.76 ± 0.77	94.00 ± 0.11
MSDC-Swin-T	83.74 ± 0.63	90.50 ± 1.57	94.38 ± 1.13	96.50 ± 0.47

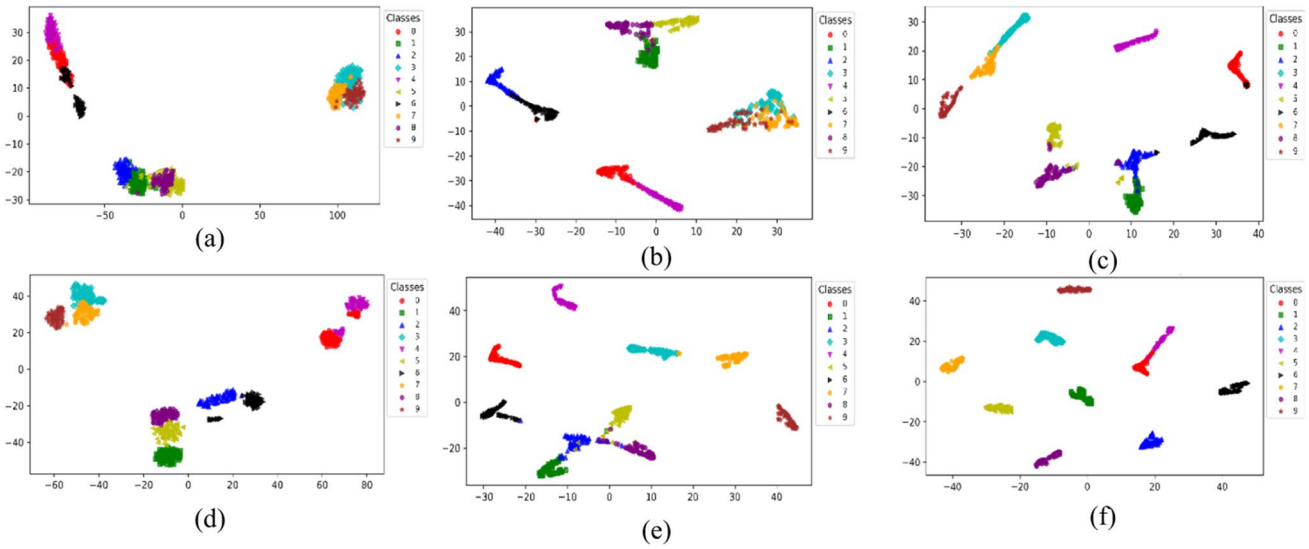


Fig. 12 Visualization results of the six methods at SNR = 0 dB in Case 1. (a) Swin-T. (b) EAPN. (c) RWKDCAE. (d) GMAML. (e) MBSDCN. (f) MSDC-Swin-T

Table 4 Diagnostic accuracy at different sample sizes

Algorithms	Amount of training samples / Amount of test samples—Accuracy (%)			
	170 / 50	350 / 100	700 / 200	1400/ 400
Swin-T	51.85 ± 3.71	57.00 ± 2.48	78.67 ± 1.07	89.25 ± 0.75
EAPN	57.41 ± 1.85	76.86 ± 0.96	86.49 ± 1.76	95.52 ± 0.44
RWKDCAE	60.21 ± 1.11	82.28 ± 1.07	88.99 ± 1.44	96.50 ± 0.17
GMAML	68.52 ± 1.14	85.21 ± 0.33	91.42 ± 0.83	95.97 ± 1.02
MBSDCN	74.07 ± 0.63	88.42 ± 0.71	94.61 ± 0.11	96.00 ± 0.31
MSDC-Swin-T	77.78 ± 1.85	96.47 ± 0.53	98.10 ± 0.23	99.61 ± 0.17

4.1.3 Test performance under mixed faults

To analyze the fault diagnosis performance of the proposed method under mixed faults, we use confusion matrix to obtain the classification results of Swin-T, EAPN, RWKDCAE, GMAML, MBSDCN and MSDC-Swin-T, as shown in Fig. 9. The horizontal direction represents the predicted label and the vertical direction represents the true labels, the diagonal lines represent the number of correct classifications for different fault types. In Fig. 9, the proposed method has high diagnostic accuracy for Ball, Inner, Health_1, Chipped, Surface and Health_2. For Outer fault, the fault accuracy is slightly lower than EAPN and MBSDCN. However, the proposed method has the highest overall test accuracy of 98.1%, and the other five methods have test accuracies of 78.67% (Swin-T), 86.49% (EAPN), 88.99% (RWKDCAE), 91.52% (GMAML) and 94.61% (MBSDCN). In addition, it can be observed from Fig. 9 that the five compared methods have misclassification on Chipped classification, due to the similarity between the fault types of bearings and gearboxes, resulting in a lower identification accuracy. However, the proposed method not only accurately identifies the Chipped fault category, but also has the lowest misclassification rate on the other fault types, because the designed multiscale convolutional token embedding module is able to extract fault information at different time scales. The fault information is further positionally

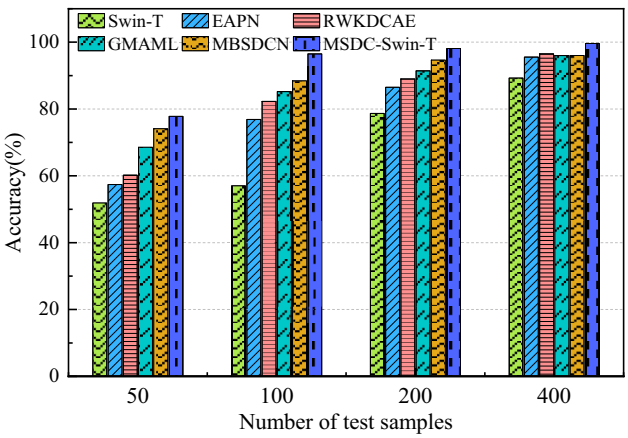


Fig. 13 Fault diagnosis results under different test samples

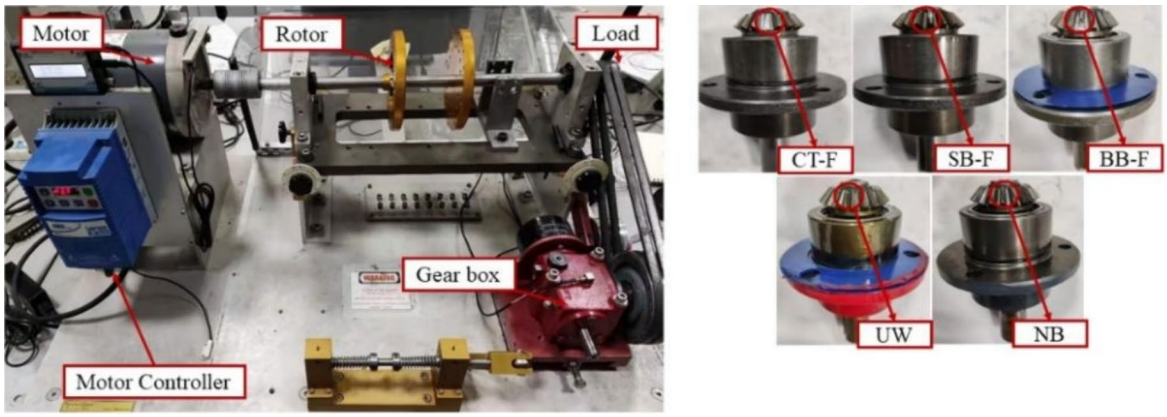


Fig. 14 Roller Gear Failure Simulation Test Platform

Table 5 Description of fault types in the dataset

Location	Type	Description	Label
Gearbox	NB	Normal bevel gear	0
	SB-T	Small-end broken half-tooth	1
	CT-F	Complete tooth fracture	2
	BB-T	Big-end broken half-tooth	3
	UW	Uniform wear	4

encoded in Swin-Transformer, which improves the accuracy of fault identification. In summary, the above analyses show that the proposed method has significant advantages in mixed faults and can effectively diagnose mixed faults.

The diagnostic accuracy and loss of the MSDC-Swin-T are shown in Fig. 10. In the experiment, the number of samples per small batch is set to 64, and the epochs are 100. In Fig. 10, the curve of the loss fluctuates in the early stages of the training process due to the randomness of the initialization parameters. With the training proceeds, the model tends to converge after the 60th epoch. The above analysis shows that the proposed method has good generalization performance, which is attributed to the use of Radam and gradient cropping.

4.1.4 Test performance with additional noise

In real engineering environments, the collected vibration signals are usually affected by noise, which makes their characterization information more complicated. Therefore, fault diagnosis in the strong noise environments is a huge challenge. In this paper, we add signal-to-noise ratio (SNR) to the original vibration signals which better simulates the real noise environments. Figure 11 shows the results of adding different SNR values to the original signals. SNR is defined as:

$$SNR = 10\log_{10}(P_{singal}/P_{noise}) \tag{27}$$

where P_{singal} is the power of original signals and P_{noise} is the power of noise signals.

Table 3 shows the diagnostic results of the six methods in noisy environment, by analyzing the results, we find that the diagnostic accuracy of each method decreases accordingly as SNR reduces continuously. When SNR is 4 and 8 dB, the diagnostic accuracies of MSDC-Swin-T are 94.38% and 96.5% respectively, which are higher than the other comparison methods. This indicates that normal noise has limited influence for the proposed method. In contrast, the comparison methods are more susceptible to noise disturbances. Especially for the Swin-T method, the diagnostic accuracy is only 74.54% at SNR = 8 dB. When SNR is -4 dB, the diagnostic accuracy of Swin-T is only 55.92%, the diagnostic accuracy of MSDC-Swin-T is 83.74%. This is attributed to the fact that the designed multiscale convolutional token embedding module performs feature information extraction at different scales, which can mine sufficient low-order local feature information. In addition, the Swin-Transformer with Sparse Tokens module further positionally encodes the low-order local features to capture the periodic fault information in the vibration signals, which improves the fault identification accuracy of the model.

To further observe the feature extraction performance of MSDC-Swin-T, we use t-SNE to map high-dimensional features into a two-dimensional plane, as shown in Fig. 12. In Fig. 12(a) and (b), the features of the original signals are mixed together in the low-dimensional space and cannot distinguish their fault types after Swin-T and EAPN feature extraction. This indicates that the feature extraction capability of Swin-T and EAPN is insufficient to extract enough fault information, resulting in low fault identification accuracy. In Fig. 12(c), (d) and (e), although the faults of different types have clear classification boundaries, there are still a considerable number of fault samples overlapping after RWKDCAE, GMAML and MBSDCN feature extraction. This indicates that RWKDCAE, GMAML and MBSDCN have certain immunity to noise, but the lack of sufficient training samples

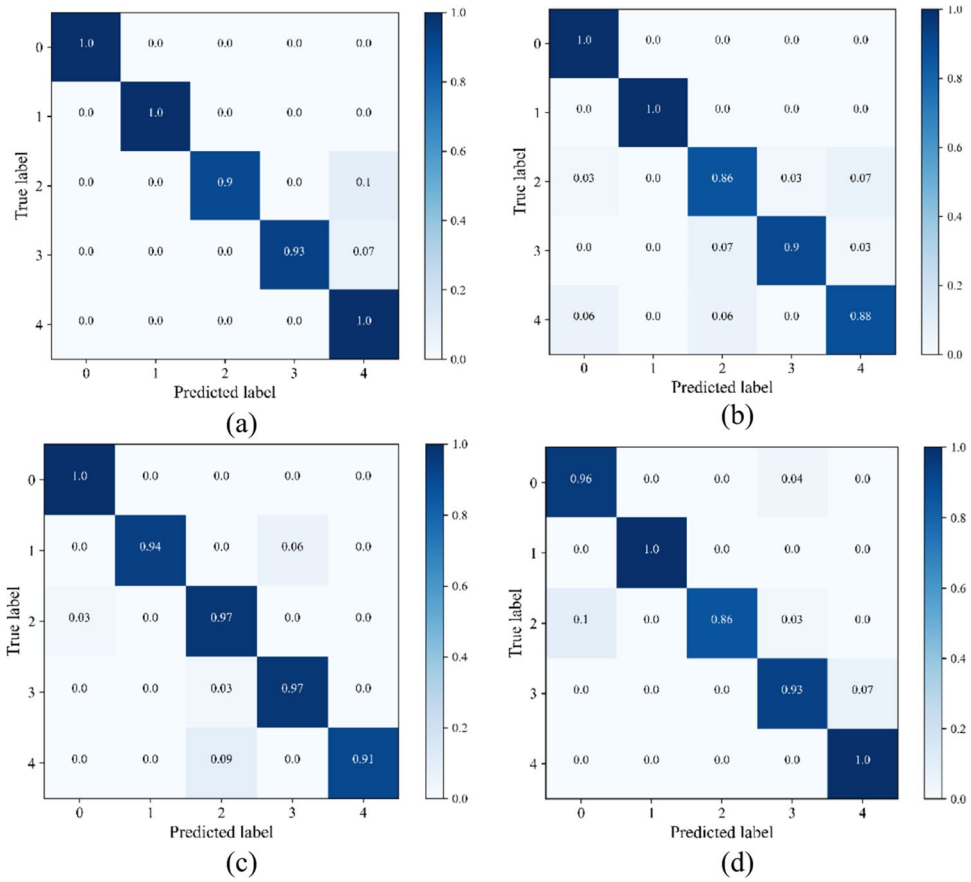


Fig. 15 Diagnostic results of the proposed method under different operating conditions. (a) 0HP. (b) 1HP. (c) 2HP. (d) 3HP

under smalls ample conditions results in that the models are unable to capture the complex feature relationships hidden in the data, which reduce the diagnostic accuracy. In Fig. 12(f), MSDC-Swin-T obtains the best clustering effect, which can better identify different types of fault samples. This means that MSDC-Swin-T can adequately extract the fault information in the data under small sample and noise environment, which improves the model's ability to capture complex data features, thus improves the accuracy of fault identification and robustness to noise interference.

Table 6 Mixed fault dataset segmentation

Location	Type	Description	Label
0HP	CT-F	Complete tooth fracture	0
	BB-T	Big-end broken half-tooth	1
1HP	CT-F	Complete tooth fracture	2
	BB-T	Big-end broken half-tooth	3
2HP	CT-F	Complete tooth fracture	4
	BB-T	Big-end broken half-tooth	5
3HP	CT-F	Complete tooth fracture	6
	BB-T	Big-end broken half-tooth	7

4.1.5 The performance with different sample sizes

In this section, the small sample conditions include 250, 500, 1000 and 2000 total sample sizes, the dataset contains 10 classes of fault types and is divided into the training, testing and validation sets according to 7:2:1. The experimental results of different methods are shown in Table 4 and Fig. 13. Figure 13 shows that MSDC-Swin-T achieves the best diagnosis accuracy under four sample sizes. When the sample size of the testing is 50, the diagnostic accuracy of MSDC-Swin-T is higher than Swin-T, EAPN, RWKDCAE, GMAML, and MBSDCN by 25.93%, 20.37%, 17.57%, 9.26%, and 3.71%, respectively. When the number of samples in the testing is 100, the diagnostic accuracy of the MSDC-Swin-T is 96.47%. However, when the number of samples in the testing is 400, the diagnostic accuracies of the compared methods can only reach about 96%. This indicates that the diagnostic performance of MSDC-Swin-T is better than other methods under small sample.

4.2 Case 2

4.2.1 Description of the dataset

This case uses the roller gear fault simulation test platform, which is shown in Fig. 14. The platform mainly

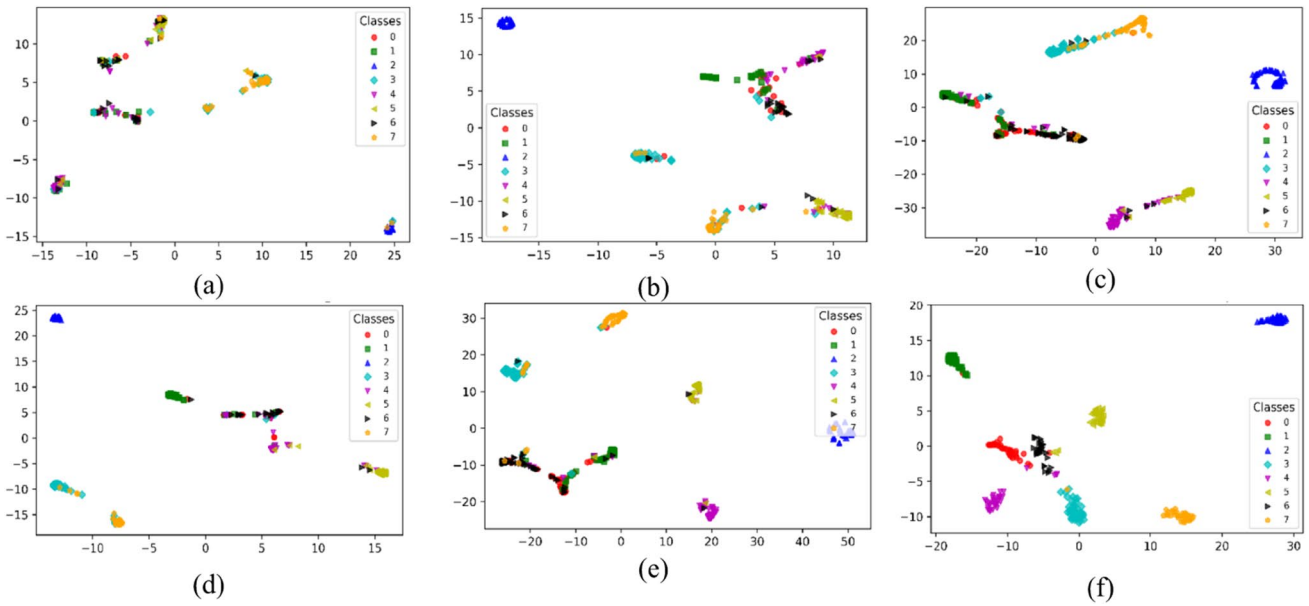


Fig. 16 Visualization of mixed fault identification results for different operating conditions. (a) Swin-T. (b) EAPN. (c) RWKDCAE. (d) GMAML. (e) MBSDCN. (f) MSDC-Swin-T

consists of a three-phase AC asynchronous motor, a loading device, a gearbox, bearings of different fault types, sensors and frequency controllers. Among them, the sensors are shear acceleration sensors, the type is 333B30 and the sensitivity is 102.7 mV/g. The frequency of the sample acquisition is 2.56 kHz. Compared with the test rig in Case 1, the test object of this test rig is bevel gears, and the data are obtained by adjusting the frequency controller and the loading device at 20 Hz for 0 HP, 1 HP, 2 HP, and 3 HP, respectively. Among them, the failure states of the bevel gears are complete tooth fracture faults, small-end broken half-tooth faults, big-end broken half-tooth faults, uniform wear faults and normal bevel gear as shown in Fig. 14. The detailed division of the dataset is shown in Table 5.

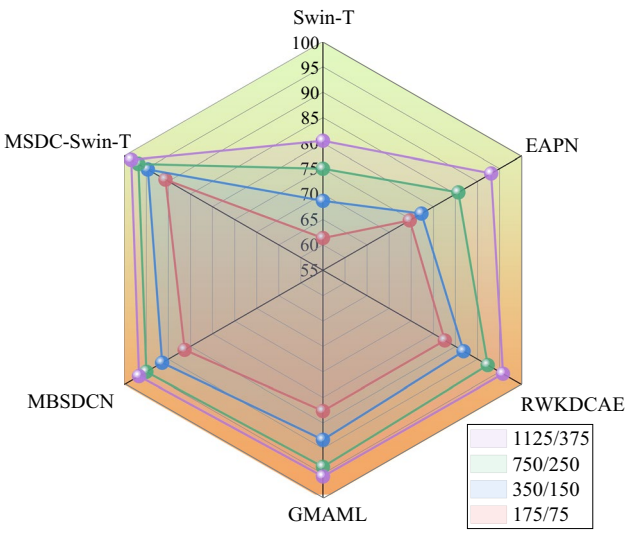


Fig. 17 Fault diagnosis results with different test samples

4.2.2 Comparison of performance under different operating conditions

According to the optimal network structure and hyperparameter settings chosen in Case 1, we train the MSDC-Swin-T model under four different working conditions in Case 2 respectively. Among them, the samples of each type fault are 100, and the total number of samples is 500. The confusion matrix of the diagnosis results is shown in Fig. 15, the proposed method has higher recognition rates for labels 0, 1 and 4. It has the disadvantage of high misclassification rate for both labels 2 and 3. The reason is that the fault characteristics of labels 2 and 3 are relatively similar, resulting in unsatisfactory results in the recognition process. However, in terms of total fault identification, the proposed method has high diagnostic accuracies for the four operating conditions, which are 92.66%, 96.59%, 95.62% and 95.07%, respectively.

4.2.3 Comparison of mixed fault performance under different operating conditions

Based on the research in the previous section, we select fault type label 2 and label 3 with the lowest accuracy under four different working conditions to construct a

Table 7 Diagnostic accuracies at different sample sizes

Algorithms	Amount of training samples / Amount of test samples—Accuracy (%)			
	175 / 75	350/ 150	750 / 250	1125/ 375
Swin-T	61.33 ± 2.67	68.67 ± 1.34	75.00 ± 0.33	80.44 ± 0.63
EAPN	74.67 ± 1.33	77.33 ± 0.67	85.67 ± 1.76	93.11 ± 1.12
RWKDCAE	82.62 ± 0.31	86.87 ± 1.13	92.33 ± 0.66	95.78 ± 0.44
GMAML	82.67 ± 1.13	88.33 ± 0.33	93.67 ± 0.13	95.56 ± 0.23
MBSDCN	86.34 ± 0.63	91.42 ± 1.25	95.00 ± 0.67	96.67 ± 0.31
MSDC-Swin-T	90.67 ± 1.35	94.67 ± 1.34	96.81 ± 0.23	98.44 ± 0.26

Table 8 Diagnostic accuracies of different methods in variable load environments

Algorithms	Different working conditions				Average
	0HP	1HP	2HP	3HP	
Swin-T	64.16%	63.33%	61.66%	62.78%	62.98%
MS-Swin-T	85.55%	90.00%	88.44%	88.15%	88.03%
CNN-Swin-T	82.22%	88.00%	86.60%	85.00%	85.45%
WSwin-T	86.67%	92.74%	92.22%	90.55%	90.54%
MSDC-Swin-T	91.45%	96.11%	95.55%	94.44%	94.39%

new dataset, which is shown in Table 6, this dataset is used to validate the mixed fault diagnosis performance under different operating conditions. Where each fault type is 100 samples and the total number of samples is 800. To intuitively understand the quality of feature extraction, t-SNE is used for downscaling and visual features, the results are shown in Fig. 16. In Fig. 16(f), the proposed method has more discriminative clustering effect for the eight fault types, and the different fault types are separated independently with large interclass distances. For Swin-T, EAPN and RWKDCAE, most of the fault types are mixed and the fault identification performance is poor. For GMAML and MBSDCN, although there are distinct aggregation boundaries for some fault types, a few fault types are mixed. In addition, the proximity between the classes of different fault types leads to a higher risk of misclassification.

4.2.4 Comparison of performance with different sample sizes

The diagnostic results under different sample sizes are shown in Fig. 17 and Table 7. In Fig. 17, we can more intuitively see that the diagnostic accuracies of the proposed method are higher than that of the other comparative methods under all four small sample conditions. The diagnostic accuracy of MSDC-Swin-T remains above 90% as the number of test samples decreases from 375 to 75, which indicates that the proposed method can accurately identify fault categories under more stringent small sample conditions. In Table 7, we can see that the accuracies of Swin-T, EAPN, RWKDCAE, GMAML, and MBSDCN are 61.33%, 74.67%, 82.62%,

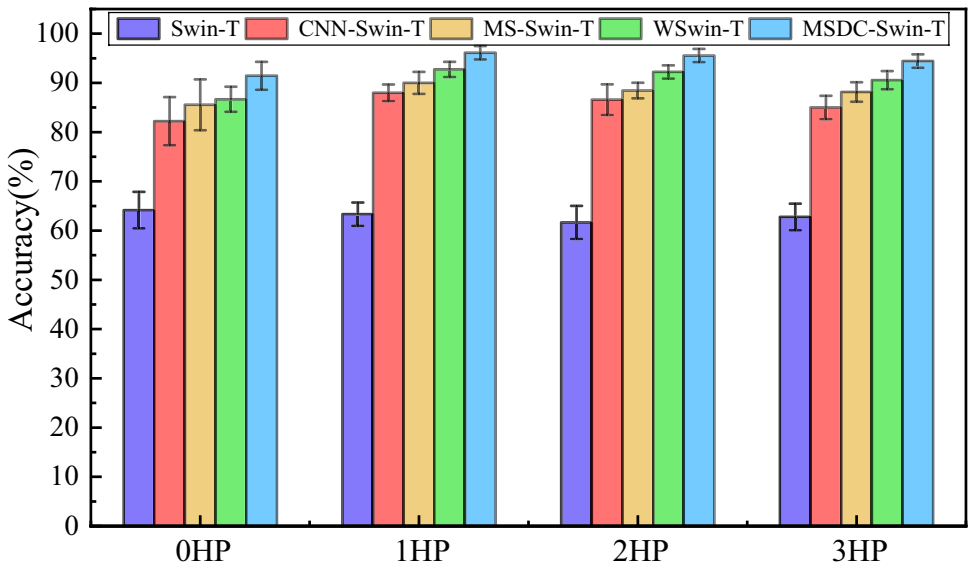


Fig. 18 Diagnostic accuracy of different methods in variable load environment

82.67%, and 86.34% for 75 test samples, while the diagnostic accuracy of the MSDC-Swin-T is 90.67%. In addition, the diagnostic accuracy of the MSDC-Swin-T gradually increases as the number of test samples increases. When the test samples are 375, the diagnostic accuracy of the proposed method reaches the highest of 98.44%. This shows that MSDC-Swin-T can effectively mine feature information.

4.2.5 Ablation experiments

In order to explore the effectiveness of each module of the proposed method, we conduct ablation experiments under different working conditions in Case 2. Specifically, the proposed method is the improvement of the Swin-Transformer model. The improvements include the utilization of multiscale feature extraction technique, the design of coordinate reconstruction attention mechanism, and token sparse operation. Therefore, we establish four comparison methods, namely: without multiscale convolutional token embedding module (Swin-T), without.

coordinate reconstruction attention (MS-Swin-T), without multiscale feature extraction module (CNN-Swin-T), and without token sparse operation (WSwin-T). The hyperparameters of all compared methods are consistent with the proposed method. The diagnostic results are shown in Table 8 and Fig. 18.

The following conclusions can be concluded from Table 8 and Fig. 18. (1) By comparing MS-Swin-T and CNN-Swin-T with Swin-T, it is shown that low-order local feature extraction can greatly improve the diagnostic accuracy. Specifically, compared with Swin-T, the average diagnostic accuracies of MS-Swin-T and CNN-Swin-T are improved by 25.05% and 22.47% under four working conditions, respectively. (2) The average diagnostic accuracy of MS-Swin-T is improved by 2.58% compared with CNN-Swin-T. This indicates that after using the multi-scale feature extraction module, the fault features under different time scales can be effectively extracted to improve the diagnosis accuracy. (3) The average diagnostic accuracy of WSwin-T is improved by 2.51% compared with MS-Swin-T. This indicates that our designed coordinate reconstruction attention can focus on more periodic fault.

features and suppress interference information under variable load environments. (4) The average diagnosis accuracy of MSDC-Swin-T is improved by 3.85% compared with WSwin-T. This indicates that the token sparse operation makes the model training more stable and helps the model to improve the diagnosis performance.

5 Conclusion

In this paper, a multiscale dilated convolution and Swin-Transformer (MSDC-Swin-T) method is proposed for small sample fault diagnosis of gearboxes. MSDC-Swin-T solves the problem that the diagnostic model is subject to environmental noise, variable load environments. and insufficient fault samples leading to performance degradation. Specifically, we first design a multiscale convolutional embedding module for extracting low-order local features. Then, the feature map is decomposed into a series of tokens by a segmentation technique, and the self-attention in Swin-Transformer is used for modeling global dependencies. Finally, MSDC-Swin-T is validated by two gearbox datasets, and the experimental results show that MSDC-Swin-T obtains better fault recognition accuracy than existing state-of-the-art methods.

In future, we will study the applications of unsupervised learning and Swin-Transformer to further reduce the dependence on labeled samples, and to improve the performance of gearbox fault diagnosis under small sample.

Acknowledgements This work is supported by the National Natural Science Foundation of China (No.62263021), the College Industrial Support Project of Gansu Province (2023CYZC-24), the Youth Science and Technology Fund Program of Gansu Province (No.22JR5RA808)

Data availability Data available on request from the authors.

Declarations

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.

References

1. Ye Z, Yu J (2022) Deep negative correlation multisource domains adaptation network for machinery fault diagnosis under different working conditions. IEEE/ASME Trans Mechatron 27:5914–5925
2. Xie X, Chen W, Chen B et al (2020) Comprehensive fatigue estimation and fault diagnosis based on Refined Generalized Multi-Scale Entropy method of centrifugal fan blades. Measurement 166:108224
3. Wang B, Zhang M, Xu H et al (2023) A cross-domain intelligent fault diagnosis method based on deep subdomain adaptation for few-shot fault diagnosis. Appl Intell 53:24474–24491
4. Jiang G, Xie P, He H et al (2017) Wind turbine fault detection using a denoising autoencoder with temporal information. IEEE/ASME Trans Mechatron 23:89–100
5. Zhao X, Zhang Y (2022) An intelligent diagnosis method of rolling bearing based on multi-scale residual shrinkage convolutional neural network. Meas Sci Technol 33:085103
6. Liang H, Cao J, Zhao X (2022) Multi-scale dynamic adaptive residual network for fault diagnosis. Measurement 188:110397
7. Wei A, Han S, Li W et al (2023) A new framework for intelligent fault diagnosis of spiral bevel gears with unbalanced data. Appl Intell 53:21312–21324
8. Jie D, Zheng G, Zhang Y et al (2021) Spectral kurtosis based on evolutionary digital filter in the application of rolling element bearing fault diagnosis. Int J Hydrol 4:27–42
9. Chen B, Song D, Cheng Y et al. (2022) IGIgram: An improved Gini index-based envelope analysis for rolling bearing fault diagnosis. J Dyn Monitoring Diag 111–124

10. Zhao M, Zhong S, Fu X et al (2020) Deep residual networks with adaptively parametric rectifier linear units for fault diagnosis. IEEE Trans Industr Electron 68:2587–2597
11. Zhang K, Chen J, Zhang T et al (2020) Intelligent fault diagnosis of mechanical equipment under varying working condition via iterative matching network augmented with selective signal reuse strategy. J Manuf Syst 57:400–415
12. Feng Y, Chen J, Yang Z et al (2021) Similarity-based meta-learning network with adversarial domain adaptation for cross-domain fault identification. Knowl-Based Syst 217:106829
13. Ma Y, Jiao L, Liu F et al (2023) Curvature-balanced feature manifold learning for long-tailed classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 15824–15835
14. Xiang W, Liu S, Li H et al (2023) Fault diagnosis of gearbox based on refined topology and spatio-temporal graph convolutional networks. IEEE Sens J 24:1866–1879
15. Ma R, Han T, Lei W (2023) Cross-domain meta learning fault diagnosis based on multi-scale dilated convolution and adaptive relation module. Knowl-Based Syst 261:110175
16. Zhang X, Huang W, Wang R et al (2023) Multi-stage distribution correction: A promising data augmentation method for few-shot fault diagnosis. Eng Appl Artif Intell 123:106477
17. Zhang K, Chen Q, Chen J et al (2022) A multi-module generative adversarial network augmented with adaptive decoupling strategy for intelligent fault diagnosis of machines with small sample. Knowl-Based Syst 239:107980
18. Lin J, Shao H, Zhou X et al (2023) Generalized MAML for few-shot cross-domain fault diagnosis of bearing driven by heterogeneous signals. Expert Syst Appl 230:120696
19. Ma W, Zhang Y, Ma L et al (2023) An unsupervised domain adaptation approach with enhanced transferability and discriminability for bearing fault diagnosis under few-shot samples. Expert Syst Appl 225:120084
20. Zheng X, Yue C, Wei J et al (2023) Few-shot intelligent fault diagnosis based on an improved meta-relation network. Appl Intell 53:30080–30096
21. Shi P, Wu S, Xu X et al (2023) TSN: A novel intelligent fault diagnosis method for bearing with small samples under variable working conditions. Reliab Eng Syst Saf 240:109575
22. 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
23. Xiao Y, Shao H, Wang J et al (2024) Bayesian Variational Transformer: A generalizable model for rotating machinery fault diagnosis. Mech Syst Signal Process 207:110936
24. Peng J, Shao H, Xiao Y et al (2024) Industrial surface defect detection and localization using multi-scale information focusing and enhancement GANomaly. Expert Syst Appl 238:122361
25. Ding Y, Jia M, Miao Q et al (2022) A novel time–frequency Transformer based on self–attention mechanism and its application in fault diagnosis of rolling bearings. Mech Syst Signal Process 168:108616
26. Wu H, Triebe MJ, Sutherland JW (2023) A transformer-based approach for novel fault detection and fault classification/diagnosis in manufacturing: A rotary system application. J Manuf Syst 67:439–452
27. Han S, Shao H, Cheng J et al (2022) Convformer-NSE: A novel end-to-end gearbox fault diagnosis framework under heavy noise using joint global and local information. IEEE/ASME Trans Mechatron 28:340–349
28. Li S, Ji J, Xu Y et al (2024) Dconformer: A denoising convolutional transformer with joint learning strategy for intelligent diagnosis of bearing faults. Mech Syst Signal Process 210:111142
29. Zhao X, Luo W (2023) A Deep Intelligent Hybrid Model for Fault Diagnosis of Rolling Bearing. Journal of Vibration Engineering & Technologies 11:721–737
30. Xiao Y, Shao H, Min Z et al (2022) Multiscale dilated convolutional subdomain adaptation network with attention for unsupervised fault diagnosis of rotating machinery cross operating conditions. Measurement 204:112146
31. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. Adv Neural Inf Process Syst 30:5998–6008
32. Hou Q, Zhou D, Feng J (2021) Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 13713–13722
33. Zhai X, Qiao F, Ma Y et al (2022) A novel fault diagnosis method under dynamic working conditions based on a CNN with an adaptive learning rate. IEEE Trans Instrum Meas 71:1–12
34. Wang C, Sun H, Cao X (2021) Construction of the efficient attention prototypical net based on the time–frequency characterization of vibration signals under noisy small sample. Measurement 179:109412
35. Yang D, Karimi HR, Sun K (2021) Residual wide-kernel deep convolutional auto-encoder for intelligent rotating machinery fault diagnosis with limited samples. Neural Netw 141:133–144
36. Liang H, Cao J, Zhao X (2023) Multibranch and Multiscale Dynamic Convolutional Network for Small Sample Fault Diagnosis of Rotating Machinery. IEEE Sens J 23:8973–8988

Yazhou Zhang received his M.E. degree from Lanzhou University of Technology in 2023. He is currently pursuing the Ph.D. degree with the Control Theory and Control Engineering, Lanzhou University of Technology, Lanzhou, China. His current research interests include fault diagnosis and machine learning.

Xiaoqiang Zhao received his Ph.D. degree from Zhejiang University in 2006. Now he is a professor and Ph.D. supervisor in Lanzhou University of Technology. His main research interest includes process monitoring and fault diagnosis, production scheduling, and data mining.

Haopeng Liang received his M.E. degree from Lanzhou University of Technology in 2021. He is currently pursuing the Ph.D. degree with the Department of Industrial Engineering, Lanzhou University of Technology, Lanzhou, China. His current research interests include fault diagnosis and machine learning.

Peng Chen received his Ph.D. degree from Lanzhou University of Technology in 2021. Now he is an associate professor at Lanzhou Petrochemical University of Technology. His current research interests include fault diagnosis and prediction and machine learning.

Measurement Science and Technology

PAPER

A dual-stream temporal convolutional network for remaining useful life prediction of rolling bearings

To cite this article: Yazhou Zhang *et al* 2025 *Meas. Sci. Technol.* **36** 016206

IOP Publishing Measurement Science and Technology
Meas. Sci. Technol. **36** (2025) 016206 (16pp) https://doi.org/10.1088/1361-6501/ad876e

A dual-stream temporal convolutional network for remaining useful life prediction of rolling bearings

Yazhou Zhang[1,2], **Xiaoqiang Zhao**[1,2,*], **Rongrong Xu**[1,2] **and Zhenrui Peng**[1,2]

[1] College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, People's Republic of China
[2] Gansu Key Laboratory of Advanced Control of Industrial Processes, Lanzhou 730050, People's Republic of China

E-mail: xqzhao@lut.edu.cn

Received 13 August 2024, revised 24 September 2024
Accepted for publication 16 October 2024
Published 24 October 2024

Abstract

Remaining useful life (RUL) prediction plays an indispensable role in the reliable operation and improved maintenance of rolling bearings. Currently, data-driven methods based on deep learning have made significant progress in RUL prediction. However, most of such methods only consider the correlation between channels, ignoring the importance of different time steps for RUL prediction. In addition, it is still challenging to effectively fuse the degradation features of rolling bearings to improve the model RUL prediction performance. To address the above issues, this paper proposes a novel data-driven RUL prediction method named dual-stream temporal convolution network (DSTCN). First, a hybrid attention temporal convolution block (HATCB) is designed to capture the correlation of degraded features on the channel dimension and temporal dimension. Second, a one-dimensional attention fusion module is designed. This module is capable of weight recalibration and assignment to adaptively fuse different degraded features. Afterward, the Hilbert Marginal spectrum is obtained using the Hilbert–Huang Transform and used as the input to one stream. Meanwhile, vibration signals are used as the input of the other stream, thus building a DSTCN to realize RUL prediction. The effectiveness of the proposed method is validated with two life-cycle datasets, and the results show that the method has lower prediction error than other methods for RUL prediction and prognostic analysis.

Keywords: remaining useful life prediction, rolling bearings, dual-stream temporal convolutional network, hybrid attention, feature fusion

1. Introduction

In recent years, prognostics and health management (PHM), including fault diagnosis [1], failure monitoring [2], and remaining useful life (RUL) prediction [3, 4], has become a hot topic of research in industrial applications [5–8]. Due to the different working intensity and working time of parts in mechanical systems, it leads to difficulties in accurately obtaining equipment degradation characteristics. Therefore, RUL prediction is considered as the most challenging task in PHM. Rolling bearings are widely used in rotating machinery and are known as 'industrial joints', and their operating conditions directly affect the safe operation of the whole rotating machinery [9–12]. Thus, it is of great significance to make accurate RUL predictions for rolling bearings and ensure stable operation and safe production of rotating equipment [13–15].

Currently, RUL prediction methods are categorized into physical model-based prediction methods and data-driven prediction methods [16–18]. Physical model-based prediction

* Author to whom any correspondence should be addressed.

methods require the establishment of a physical models that accurately describes the degradation of the mechanical system, include particle filtering [19], wiener process modeling [20], and eyrling modeling [21]. However, they require a large amount of prior knowledge in the construction of RUL prediction models. With the rapid development of sensor technology and storage technology, data-driven prediction methods have gradually become the mainstream in the field of PHM. The data-driven prediction methods are based on a large amount of historical data to build the models of mechanical system degradation characteristics, and their process consists of degradation indicator establishment and degradation feature learning. Classical data-driven methods, such as Support Vector Regression [22] and artificial neural networks (ANN) [23], have insufficient feature extraction capabilities, leading to poor prediction accuracies. However, with the development of deep learning, RUL prediction methods based on deep learning have received more and more attention by scholars [24].

Deep learning has powerful feature extraction capability that automatically establishes nonlinear mapping relationships between historical data and RUL prediction. In RUL prediction, convolutional neural network (CNN) [25, 26] and recurrent neural network (RNN) [27–29] are the most popular and have achieved encouraging performance. For example, Qiu *et al* [30] proposed a temporal convolutional network RUL prediction method with an adaptive degradation stage segmentation strategy, which was capable of estimating the health and degradation states in segments. Zhu *et al* [31] proposed a dynamically activated convolutional network that mined the correlations between temporal information of neighboring samples to improve prediction accuracy. Sun *et al* [32] proposed a lightweight Bi-LSTM with automatic pruning for bearing RUL prediction.

In addition, the attention mechanism can selectively focus on valuable information and suppress irrelevant information, which can improve the performance of the model. Therefore, it is widely used in RUL prediction models. For example, Zhang *et al* [33] proposed an improved CNN for high accuracy prediction, which used deep separable convolutions to construct CNNs and introduced an attention mechanism with soft thresholding to improve the network's RUL prediction performance in noisy environments. Shen *et al* [34] proposed a multi-head attention bidirectional-long-short-term-memory (MHA-BiLSTM) network for rolling bearing RUL prediction. Wang *et al* [35] proposed a new dual competitive temporal convolutional network to predict the RUL of rolling bearings, which designed a global competitive attention module to enhance the feature representation.

The above deep learning-based RUL prediction methods usevibration signals as model input. Some researchers have tried to process the vibration signals using signal processing techniques, and input them into deep learning models for RUL prediction. For example, Cao *et al* [36] used signal processing techniques to obtain marginal spectra and input them to a temporal convolutional network for feature extraction. Shi *et al* [37] used multiple stacked sparse self-encoders to learn different degradation features from different feature domains of the original data. Meanwhile, some scholars have also tried to use the processed data and vibration signals as model inputs for RUL prediction. For example, Jiang *et al* [38] proposed a convolutional two-channel transformer network with time-window cascading to extract degradation features from both time and frequency domain perspectives. Xu *et al* [39] proposed a dual-stream self-attentive neural network (DS-SANN) for RUL prediction, which used domain knowledge to obtain auxiliary data, and used the model to extract the features from vibration signals and auxiliary data.

In summary, deep learning-based RUL prediction methods have made outstanding contributions in the field of RUL prediction for rolling bearings. However, they still have some limitations: (1) RUL prediction methods using the attention mechanism usually prioritize the improvement of the model's ability to capture complex relationships across channels. However, they ignore the effect of different time steps of the input sequence on the RUL prediction performance of mechanical systems. (2) Multi-channel inputs can provide more degradation features of rolling bearings. However, how to fuse these degradation features can also significantly affect the final RUL prediction results. To address the above issues, this paper proposes a dual-stream temporal convolutional network (DSTCN) for the RUL prediction of rolling bearings. Specifically, firstly, the designed hybrid attention temporal convolution block (HATCB) is not only capable of capturing complex relationships across channels, but also pay more attention to degraded information in the temporal dimension. Secondly, a one-dimensional attention fusion module is designed, which can adaptively select important degraded features of different streams to be fused, so as to avoid redundant information caused by adding different stream features. Finally, the Hilbert Marginal spectrum is obtained using the Hilbert–Huang Transform, which is used as the input of one stream. Meanwhile, vibration signals are used as the input of the other stream, thus building a dual-stream temporal convolution network (DSTCN) to realize RUL prediction. The main contributions of this paper are summarized as:

(1) A HATCB is designed to capture the correlation between features on the channel dimension and the temporal dimension. Specifically, a parallel hybrid attention mechanism is constructed using channel attention and temporal attention to achieve the feature weighting on the channel dimension and the temporal dimension.
(2) A one-dimensional attention fusion module is designed. The extracted feature information from different streams is adaptively fused to avoid redundant information caused by direct addition of feature information.
(3) Hilbert Marginal spectra (HMs) and vibration signals are used as the inputs to different streams, and HATCB is used as a basic component to construct the DSTCN to realize RUL prediction.

The rest of the paper is organized as follows. Section 2 introduces the related basic theories. Section 3 introduces the proposed prediction method. Section 4 presents the

effectiveness and superiority of the proposed method. Section 5 provides the conclusion.

2. Preliminaries

2.1. Marginal spectrum calculation

Frequency domain signals are obtained by Fourier transform, which can provide frequency information in the global range and cannot obtain frequency information at different time points. Unlike Fourier transform, HMs have a significant advantage in dealing with non-smooth signals [40]. It extracts the instantaneous frequency and amplitude information at different time points in the signals. Therefore, to avoid the missing of degradation information, HMs are used as the input to one of DSTCN. HMs are obtained by Hilbert–Huang Transform. Specifically, firstly, the vibration signals are decomposed by EMD into a set of intrinsic modal functions with different time scales. The process is described as follows:

$$f(t)=\sum_{i=1}^{n}x_i(t)+r_n(t) \tag{1}$$

where n is the number of intrinsic modal functions, $x_i(t)$ is the eth intrinsic modal function, and $r_n(t)$ is the final residual term. Then, Hilbert–Huang Transform is applied to each intrinsic modal function, and the process is as follows:

$$x_i^{\mathrm{H}}(t)=\frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{x_i(\tau)}{t-\tau}\mathrm{d}\tau. \tag{2}$$

Secondly, the analytic signal of $x_i(t)$ is constructed as follows:

$$z_i(t)=x_i(t)+jx_i^{\mathrm{H}}(t)=a_i(t)\,\mathrm{e}^{j\varphi_i(t)} \tag{3}$$

where $a_i(t)=\sqrt{x_i(t)^2+x_i^{\mathrm{H}}(t)^2}$ is the instantaneous amplitude and $\varphi_i(t)=\arctan\left[x_i^{\mathrm{H}}(t)/x_i(t)\right]$ is the instantaneous phase. Further the instantaneous frequency can be expressed as follows:

$$h(\omega)=\int_0^T x_i^{\mathrm{H}}(\omega,t)\,\mathrm{d}t\ \ \omega_i(t)=\frac{\mathrm{d}\varphi_i(t)}{\mathrm{d}t} \tag{4}$$

where T denotes the end time of the time–frequency distribution function. Finally, HMs $h(\omega)$ are obtained by integrating the time–frequency distribution function $\omega_i(t)$. The process is as follows:

$$h(\omega)=\int_0^T x_i^{\mathrm{H}}(\omega,t)\,\mathrm{d}t. \tag{5}$$

HMs provide the total amplitude contribution of each frequency value. Therefore, it is more suitable for non-smooth signal analysis.

2.2. Dilated causal convolution

Dilated causal convolution consists of two parts: dilated convolution and causal convolution. The output information of dilated convolution is exponentially related to the number of network layers. Thus, dilated convolution helps the network to obtain a larger receptive field. The output of causal convolution at moment t is only associated with the output element at an earlier moment ($\{t-1,t-2,\ldots,1\}$), which can ensure that the historical data is not leaked. Thus, dilated causal convolution can be defined as follows:

$$F_i=(x_d*k)_i=\sum_{f=0}^{c-1}x_{i+f\bullet d}\cdot k_{i+f}+b \tag{6}$$

where $*$ denotes a convolution operation, d denotes the dilated factor, k_{i+f} denotes the convolution kernel, c denotes the number of convolution kernels, $x_{i+f\cdot d}$ denotes the elements of the input sequence, and b denotes the convolution kernel bias. When the dilated factor $d=1$, dilated causal convolution is the standard convolution.

2.3. Kalman filter

Kalman filter is an efficient optimal estimator, which is proposed by R. E. Kalman and R. S. Bucy in 1961. It is widely used for data noise reduction [41], and its main computational procedure is as follows.

(1) The optimal result from the previous moment is used to obtain the predicted value of the current moment, and the process can be described as follows:

$$\overleftarrow{x}_t=F\hat{x}_{t-1}+Bu_{t-1} \tag{7}$$

where F is the state transfer matrix, B is the control matrix, $\hat{x}_{t-1}$ is the state value at the previous moment, and u_{t-1} is the input.

(2) The observed values at the current time are used to correct the predicted values at the current time and obtain the optimal results at the current time. The process is as follows:

$$\overleftarrow{P}_t=FP_{t-1}F^T+Q \tag{8}$$

where P_{t-1} is the variance of x_{t-1}, Q is the variance of the noise signal, and F^T is the transpose matrix of the state transfer matrix.

(3) The Kalman Gain is obtained from the predicted value covariance and hyperparameters at the current moment. The process is as follows:

$$K_t=\overleftarrow{P}_tH^T\left(H\overleftarrow{P}_tH^T+R\right)^{-1} \tag{9}$$

where H is the observation matrix, H^T is the transpose matrix of the observation matrix, and R is the variance of the observation noise.

(4) The optimal estimate at the current moment is obtained from the predicted value $\overleftarrow{x}_t$, the observed value Z_t, and the Kalman Gain K_t. The process is described as follows:

$$x_t = \overleftarrow{x}_t + K_t\left(Z_t - H\overleftarrow{x}_t\right) \tag{10}$$

where Z_t is the observed value at the current moment.

(5) The covariance of the predicted value at the current moment and Kalman Gain get the covariance of the optimal estimate at the current moment. The process is described as follows:

$$P_t = (I - K_t H)\overleftarrow{P}_t \tag{11}$$

where K_t is the Kalman Gain and I is the unit matrix.

3. The proposed method

3.1. HATCB

TCN can abstract the features of the input sequence into the feature matrix, but all features in the feature matrix do not contribute equally to the RUL prediction. The attention mechanism can assign the weights to the features through the weight matrix, which makes the model focus on the features that are relevant to the current task and suppress the irrelevant features. Therefore, we embed the attention mechanism into TCN to construct a HATCB, as shown in figure 1.

In figure 1, HATCB consists of dilated causal convolution, batch normalization, activation function, and hybrid attention. Dilated causal convolution not only ensures that historical data are not forgotten, but also has a large receptive field that can help the network capture more degraded information. Hybrid attention consists of channel attention and temporal attention. It assigns the weights to the features in the feature matrix from the channel dimension and temporal dimension, respectively. Thus, the model pays more attention to the features that are valuable to RUL and suppresses unnecessary features.

The specific process of HATCB is as follows.

Step1: the input feature matrix is firstly extracted by dilated causal convolution, and the features are processed by batch normalization function and activation function to improve the expression performance of the model. The process can be described as follows:

$$Z = F_{\mathrm{Dconv}}(X) = \sum_{i,j=0}^{c=0} x_{i,j}^{c} \cdot k + b \tag{12}$$

$$\hat{Z} = \mathrm{BN}\left(\mathrm{LRelu}(Z)\right) \tag{13}$$

where $F_{\mathrm{Dconv}}(\cdot)$ is the dilated causal convolutional feature extraction process, $\mathrm{BN}(\cdot)$ is the batch normalization operation, $\mathrm{LRelu}(\cdot)$ is the activation function, X is the feature matrix, $x_{i,j}^{c}$ is the feature element in the feature matrix, k is the convolutional kernel, and b is the bias.

Step2: channel features are weighted and assigned using channel attention on the channel dimension of the feature matrix. At the same time, weighted assignment of features at different time points is performed using temporal attention on the temporal dimension of the feature matrix. The process is described as follows:

$$\hat{F}_c = M_c\left(\hat{Z}\right) \otimes \hat{Z} \tag{14}$$

$$\hat{F}_t = M_t\left(\hat{Z}\right) \otimes \hat{Z} \tag{15}$$

where $M_c \in \{i \times c\}$ is the attention weight of the channel dimension and $M_t \in \{i \times j\}$ is the attention weight of the temporal dimension. M_c is obtained by channel attention calculation, the process is described as follows:

$$M_c = \mathrm{Softmax}\left(\mathrm{Conv}\left(\mathrm{GAP}\left(\hat{Z}\right)\right)\right). \tag{16}$$

Specifically, the feature matrix is first processed by two 1×1 convolution kernels to obtain the feature maps p, q, where p, q satisfy $\{p,q\} \in R^{1\times N}$. Subsequently, the feature map q is used by the Softmax function to obtain the positional importance weights. The process is described as follows:

$$q' = \frac{\exp(q_i)}{\sum_{i=1}^{N}\exp(q_i)} \tag{17}$$

where q' is the obtained attention feature map, and q_i is the ith feature element of the feature map q.

Finally, a dot product operation is performed between the feature maps p, q'. The process is described as follows:

$$\mathrm{att} = \{p_i \cdot {q'}_i, i \in [1,N]\} \tag{18}$$

where p' is the ith feature element in the feature map p, and q_i' is the ith feature element in the feature map q'.

Step3: the features obtained from the channel dimension and the temporal dimension are summed to obtain the output feature matrix of HATCB. The process is as follows:

$$\hat{Z}' = \hat{F}_c + \hat{F}_t. \tag{19}$$

3.2. Fusion strategies

Generally, the degraded features extracted from different streams have redundancy and correlation. If the features of each stream are fused directly at the fully connected layer, it would not only increase the computational burden of the network, but also affect the RUL prediction performance of the model. Li *et al* [42] studied the attention mechanism of convolutional kernels, adaptively adjusting the receptive field size of the network by using convolutional kernel inputs with different weights. Inspired by their study, this paper designs a one-dimensional attention fusion module, as shown in figure 2. This module selects more critical information for fusion through weight allocation, which not only saves computational resources but also improves RUL prediction performance. In figure 2, assumed that F_1, F_2 denote two stream input features, the specific process of one-dimensional attention fusion module is as follows:

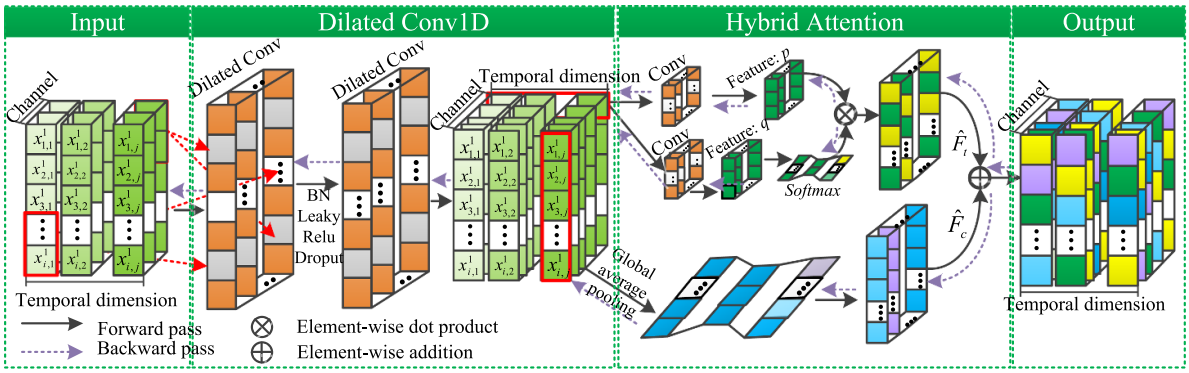


Figure 1. Hybrid attention temporal convolution block.

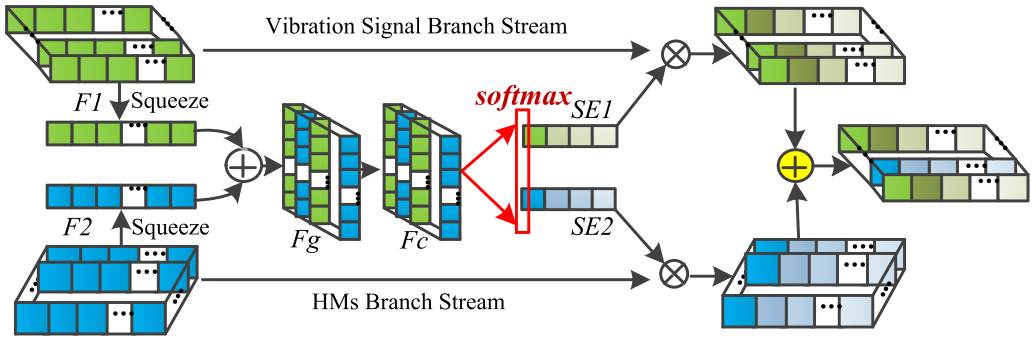


Figure 2. One-dimensional attention fusion module.

Step1: the feature information of different streams is compressed using global average pooling, the process of is described as:

$$\begin{cases} G_1^c = \frac{1}{L}\sum_{i=1}^{L} F_1^c(i) \\ G_2^c = \frac{1}{L}\sum_{i=1}^{L} F_2^c(i) \end{cases} \quad c = 1,2,\ldots,M \tag{20}$$

where G_1^c and G_2^c are the squeezed features of the cth channel, L is the spatial dimension of the feature, M is the number of channels, and $F_1^c(i)$ and $F_2^c(i)$ are the ith features of the cth channel of different streams.

Step2: the compression features of the two streams are spliced to generate the global feature F_g, and the convolution operation is performed on the feature information,

$$\begin{cases} F_g = \{G_1^c, G_2^c\} \\ F_c = f_{\mathrm{ReLU}}\left(f_{\mathrm{Norm}}\left(w \cdot F_g + b\right)\right) \end{cases} \tag{21}$$

where F_c the result of feature downscaling, w is the weight, and b is the bias. The Softmax function is used to assign the weights to global features. The process is described as follows:

$$\begin{cases} \mathrm{SE}_1 = f_{\mathrm{softmax}}\left(F_c\right) \\ \mathrm{SE}_2 = f_{\mathrm{softmax}}\left(F_c\right) \end{cases} \tag{22}$$

where $f_{\mathrm{softmax}}(\cdot)$ is the Softmax function for weight assignment.

Step3: the reassigned features and input features are multiplied to achieve recalibration and fusion of features. The process is described as follows:

$$\mathrm{F} = \mathrm{SE}_1 \otimes \mathrm{F}_1 + \mathrm{SE}_2 \otimes \mathrm{F}_2 \tag{23}$$

where F is the feature fusion and $\otimes$ is the multiply product.

3.3. DSTCN architecture

DSTCN consists of a convolutional layer, three HATCB layers, a fusion layer, and a fully connected layer, as shown in figure 3. The convolutional layer contains wide convolution and maximum pooling for suppressing noise interference in the input sequence and reducing data dimensionality. The HATCB is the core component of DSTCN. In each HATCB, we use a convolutional kernel with a convolutional scale of 3 for feature extraction, and the expansion factors are deployed according to 1, 2, and 4. In addition, to prevent the model from gradient vanishing during training, we introduce residual connectivity between each HATCB, which allows feature information to be transferred across blocks. The fusion layer refers to the adaptive weighted fusion of deep features extracted from different streams using the one-dimensional attention fusion module in section 3.2. The fully connected layer refers to the mapping of the high-dimensional features output from the fusion layer to the RUL prediction dimension. In addition, before outputting the RUL prediction results, we introduce Kalman filtering to smooth the prediction results.

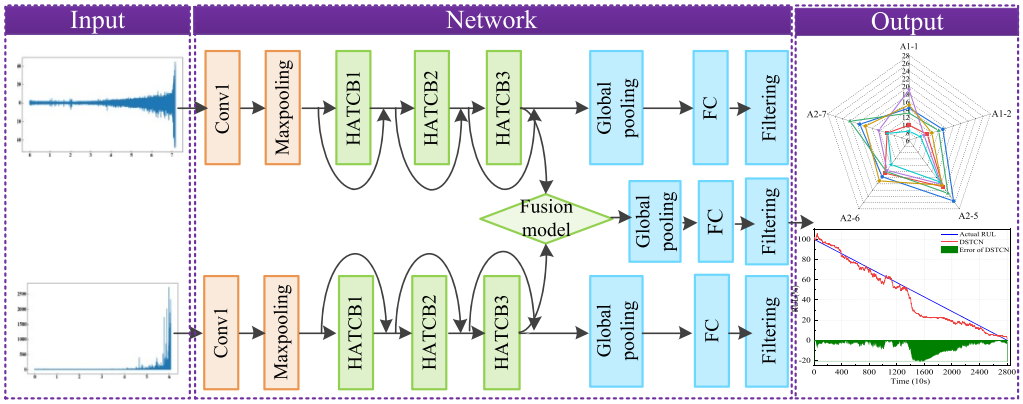


Figure 3. DSTCN architecture.

3.4. Application of DSTCN-based RUL

In this paper, a DSTCN is proposed for RUL prediction of rolling bearings. The specific steps are shown in figure 4. The process includes three stages: data preprocessing, model training, and RUL prediction. The first stage is the acquisition of raw data and data processing. The vibration signals are collected from the mechanical systems by sensors, and HMs are obtained using Hilbert–Huang Transform. The vibration signals and HMs are divided into a training set and a test set. The second stage is model construction and training. DSTCN is constructed using HATCB as the basic component. The training set is fed into the model using a moving time window. The third stage is RUL prediction. RUL prediction is performed on the test samples using the trained DSTCN.

4. Experimental results and performance analysis

4.1. Dataset description

In order to verify the prediction performance of the proposed method, two datasets are used for experimental analysis. One of them is the FEMTO dataset and the other is the XJTU-SY dataset. Figure 5 shows the experimental platform of the two datasets.

4.1.1. FEMTO dataset. The FEMTO dataset provided by the PRONOSTIA experimental platform was used to verify the effectiveness of DSTCN. The specific layout of each module of the test platform is shown in figure 5(a). The vibration signals of bearing degradation were obtained by applying radial load force. The sampling frequency was 25.6 KHz and the sampling time lasted 0.1 s. When the vibration signal amplitude exceeds 20 g, the bearing failed and the full life bearing data were obtained. The data were collected for three operating conditions, two training for each operating condition and the rest were testing. The data sets are described in detail in table 1. In this study, the vibration signals in the horizontal direction under operating conditions 1 and 2 in the FEMTO dataset are used. The training set and test set are divided as shown in table 2.

4.1.2. XJTU-SY dataset. The XJTU-SY dataset provided by Xi'an Jiao tong University was used to further validate DSTCN. Each specific layout of the test platform is shown in figure 5(b). The platform consists of a motor, a controller, a rotating shaft, a support bearing and a test bearing. The acceleration sensor type was PCB 352C33, and the vibration signals of bearing degradation were obtained by applying radial load force. The sampling frequency was 25.6 KHz, the sampling interval was 1 min. A total of 15 full-life cycle bearing degradation data were collected under three operating conditions during the test. Table 1 shows the detailed description of the data set. In this study, the vibration signals in the horizontal direction under operating conditions 1 and 2 are selected. The training set and test set are divided as shown in table 2.

4.2. Data processing

4.2.1. Calculation of the marginal spectrum. Deep learning can extract feature information directly from vibration signals. However, the degradation process of bearings has the characteristics such as nonlinearity and randomness. It is difficult to establish a following relationship between monitoring data and the degradation trend by directly using horizontal vibration signals. Hilbert Marginal spectrum can reflect the time–frequency characteristics of the rolling bearings. Therefore, one stream of the proposed method uses Hilbert Marginal spectrum as input. Taking bearing A1–1 of the FEMTO dataset as an example, the original vibration signals, frequency spectrum and edge spectrum are plotted in figure 6.

4.2.2. Data normalization. To ensure the consistency of the input of the deep learning model, the vibration signals are normalized in this paper. The process is as follows:

$$X_t^{\text{norm}} = \frac{X_t - X_t^{\text{mean}}}{X_t^{\text{std}}} \tag{24}$$

where X is the sample set $\{x_1, x_2, \ldots, x_t, x_{t+1}\}$ of horizontal vibration signals, X_t is the sample at the moment t, X_t^{mean} and X_t^{std} are the mean and standard deviation of X, respectively.

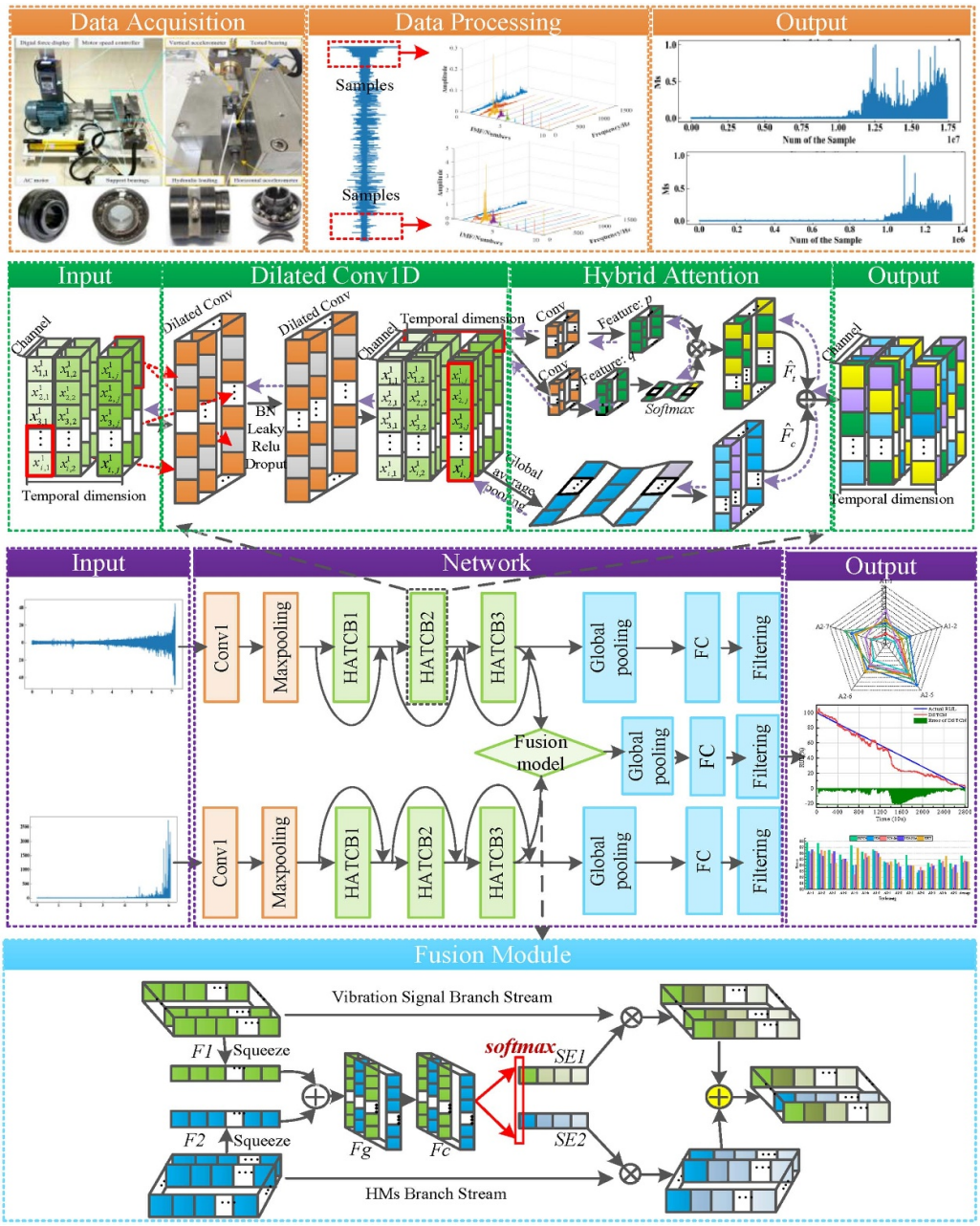
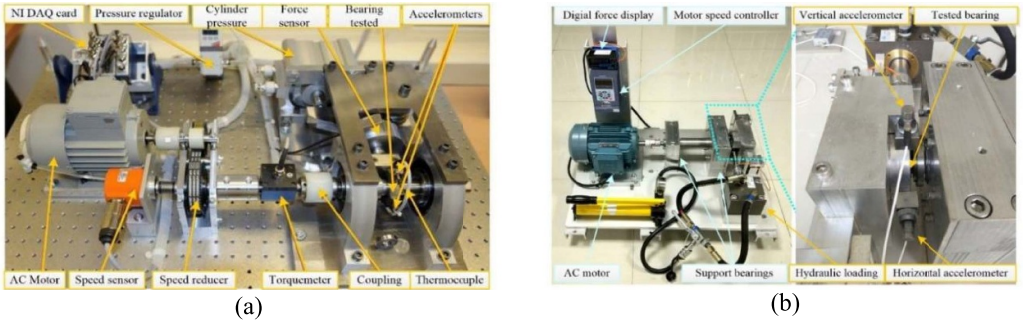


Figure 4. Flowchart of RUL based on DSTCN.



(a) (b)

Figure 5. Bearing degradation analysis test rig (a) FEMTO dataset (b) XJTU-SY dataset.

Table 1. Detailed description of FEMTO dataset and XJTU-SY dataset.

	FEMTO dataset			XJTU-SY dataset		
	Condition1	Condition2	Condition3	Condition4	Condition5	Condition6
Load (kN)	4	4.2	5	12	11	10
Speed (rmp)	4000	4200	5000	2100	2250	2400
No. Bearings	A1–1–A1–7	A2–1–A2–7	A3–1–A3–3	B1–1–B1–5	B2–1–B2–5	B3–1–B3–5

Table 2. Division of training and test sets for FEMTO dataset and XJTU-SY dataset.

FEMTO dataset					
Load (kN) 4, Speed (rmp) 4000	Training	Testing	Load (kN) 4.2, Speed (rmp) 4200	Training	Testing
	A1–2, 1–3, 1–4, 1-5, 1–6, 1–7	A1–1		A2–2, 2–3, 2–4, 2–5, 2–6, 2–7	A2–1
	A1–1, 1–3, 1–4, 1–5, 1–6, 1–7	A1–2		A2–1, 2–3, 2–4, 2–5, 2–6, 2–7	A2–2
	A1–1, 1–2, 1–4, 1–5, 1–6, 1–7	A1–3		A2–1, 2–2, 2–4, 2–5, 2-6, 2–7	A2–3
	A1–1, 1–2, 1–3, 1–5, 1–6, 1–7	A1–4		A2–1, 2–2, 2–3, 2–5, 2–6, 2–7	A2–4
	A1–1, 1–2, 1–3, 1–4, 1–6, 1–7	A1–5		A2–1, 2–2, 2–3, 2–4, 2–6, 2–7	A2–5
	A1–1, 1–2, 1–3, 1–4, 1–5, 1–7	A1–6		A2–1, 2–2, 2–3, 2–4, 2–5, 2–7	A2–6
	A1–1, 1–2, 1–3, 1–4, 1–5, 1–6	A1–7		A2–1, 2–2, 2–3, 2–4, 2–5, 2–6	A2–7
XJTU-SY dataset					
Load (kN) 12, Speed (rmp) 2100	Training	Testing	Load (kN) 11, Speed (rmp) 2250	Training	Testing
	B1–2, 1–3, 1–4, 1–5	B1–1		B2–2, 2–3, 2–4, 2–5	B2–1
	B1–1, 1–3, 1–4, 1–5	B1–2		B2–2, 2–3, 2–4, 2–5	B2–2
	B1–1, 1–2, 1–4, 1–5	B1–3		B2–1, 2–2, 2–4, 2–5	B2–3
	B1–1, 1–2, 1–3, 1–5	B1–4		B2–1, 2–2, 2–3, 2–5	B2–4
	B1–1, 1–2, 1–3, 1–4	B1–5		B2–1, 2–2, 2–3, 2–4	B2–5

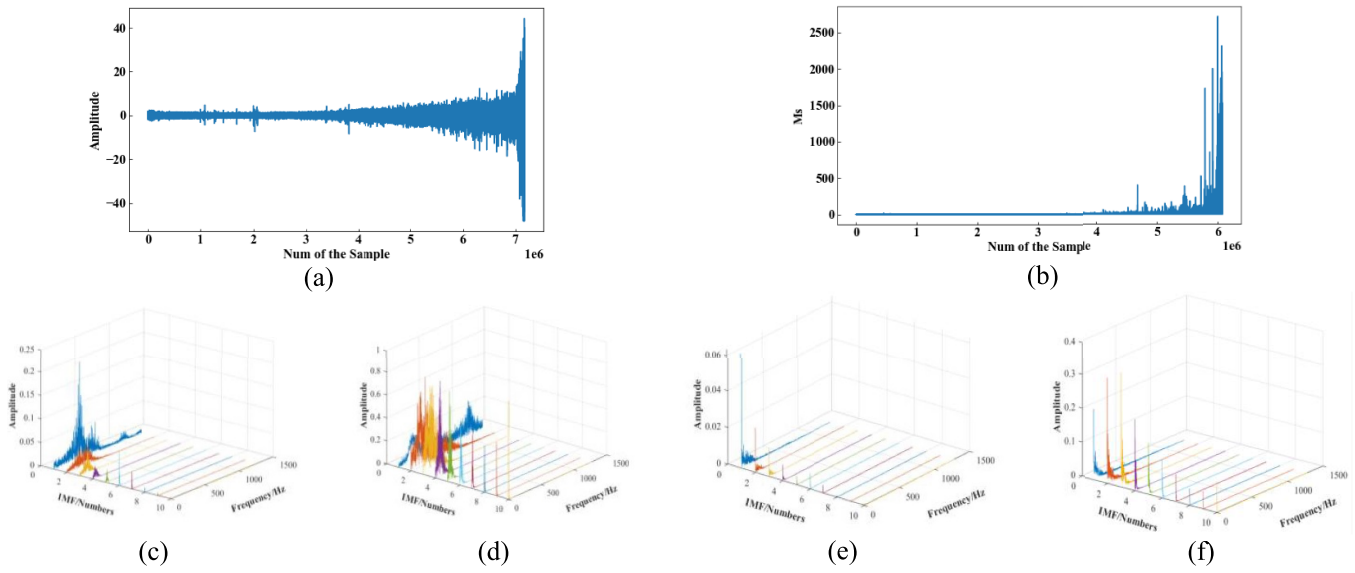


Figure 6. Bearing A1–1: (a) original signal; (b) HMs of the original signals; (c) FFT of the first sample; (d) FFT of the last sample; (e) HMs of the first sample; (f) HMs of the last sample.

Due to the different working environments, the life cycles of various devices are usually different. To enhance the model generalization, the real RUL is normalized to the RUL percentage, and the process is described as:

$$\mathrm{RUL}_t^{\mathrm{norm}} = \frac{T_{\mathrm{total-life}} - t}{T_{\mathrm{total-life}}} \tag{25}$$

where $T_{\mathrm{total-life}}$ is the total time of the device RUL, and t the time normalized value of $\mathrm{RUL}_t^{\mathrm{norm}}$ in the interval $[0 \sim 1]$.

4.2.3. Embedded time moving window. If the data from a single sample is used as input, it makes the information in the later degradation of the device irrelevant to the previous

feature information, which limits the prediction performance of the model. To solve this problem, this paper uses a time window embedding strategy to construct model input samples. The process can be described as follows:

$$X_t^{\text{input}} = \left\{x_{t-L+1}^{\text{norm}}, \ldots, x_{t-1}^{\text{norm}}, x_t^{\text{norm}}\right\} \tag{26}$$

where L is the size of the time window. The model input sample X_t^{input} consists of the current moment sample x_t^{norm} and the previous $L-1$ samples. To reduce the computational cost, a time window size of 2560 was chosen for the PRONOSTIA dataset and a time window size of 4096 was chosen for the XJTU-SY dataset.

4.3. Evaluation metrics

To compare with other advanced methods, root mean square error (RMSE), mean absolute error (MAE) and scoring function (Score) are used to evaluate the prediction performance, and the process is described as follows:

$$\text{MAE} = \frac{1}{N}\sum_{t=1}^{N}\{|\bar{y}_t - y_t|\} \tag{27}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\bar{y}_t - y_t)^2} \tag{28}$$

where N is the number of all samples, $\bar{y}_t$ is the predicted value at moment t, and y_t is the actual value at moment t.

The scoring function was used to assess the reasonableness of DSTCN prediction results. The underestimation and overestimation of the scoring function were not considered in the same way. The performance of the scoring function prefers a conservative RUL estimate [43]. The evaluation function is defined as:

$$\text{Score} = \frac{1}{N-1}\sum_{i=1}^{N-1}A_i; A_i = \begin{cases} \exp^{-\ln(0.5)\times(\text{Er}_i/5)}, \text{Er}_i \leqslant 0 \\ \exp^{\ln(0.5)\times(\text{Er}_i/20)}, \text{Er}_i > 0 \end{cases} \tag{29}$$

where N is the total number of samples and Er_i denotes the percentage of error, which can be expressed as:

$$\text{Er}_i = \frac{y_t - \bar{y}_t}{y_t} \times 100\%. \tag{30}$$

Among above evaluation metrics, the smaller the RMSE and MAE values, the better the prediction performance, and the higher the Score value, the more reasonable the prediction.

4.4. Model parameters and comparison methods

4.4.1. Model parameters. The structural configuration of DSTCN and the output size of each layer were determined through several cross-validation experiments as shown in table 3. Firstly, 1D wide convolution is used for feature extraction and a maximum pooling layer is used to reduce the number of network parameters. Secondly, HATCB consists of dilated causal convolution and hybrid attention. Dilated causal convolution can effectively capture the correlation of feature information between time series, and hybrid attention can further focus on useful information from channel dimension and temporal dimension to capture the device degradation trend. Finally, the extracted features are input to the dense layer for RUL prediction. The parameters of DSTCN are updated by Adam optimizer. The learning rate is 0.005 and the batch size is 128. The training network is 200 epoch.

Table 3. Parameter configuration of DSTCN.

Layer	Network structure parameter
Conv 1D	Kernel: 14, stride: 4, channel: 16
Activation	Leaky rate (0.1)
Maxpooling 1D	Stride: 4
HATCB1	Kernel: 3, dilated rate: 1, channel: 16, dropout rate: 0.3, leaky rate (0.1)
HATCB2	Kernel: 3, dilated rate: 2, channel: 32, dropout rate: 0.3, leaky rate (0.1)
HATCB3	Kernel: 3, dilated rate: 4, channel: 64, dropout rate: 0.3, leaky rate (0.1)
Fusion layer	Channel: 64, Activation: softmax
GAP	—
Dense	Channel: 1

4.4.2. Comparison methods. In order to fully verify the validity of DSTCN, one baseline model and three advanced prediction models are selected as comparison methods, namely TCN, TCN-SA [44], TCN-RSA [36], and CDCT [38].

(1) TCN: Consisting of dilated causal convolution, weight normalization, activation function and dense layers, the model parameters are consistent with DSTCN. The model is used to evaluate the superiority of the prediction performance between DSTCN and baseline TCN.
(2) TCN-SA: The method was proposed by Wang *et al* [44]. It consists of dilated causal convolution, soft thresholding and attention mechanism.
(3) TCN-RSA: The method was proposed by Cao *et al* [36]. It consists of dilated causal convolution and residual self-attention mechanism. This model is used to evaluate the impact of time domain branching on prediction performance.
(4) CDCT: This is the state-of-the-art model proposed by Jiang *et al* [38]. The model uses time and frequency domain signals as input and uses causal convolution for feature extraction, and ultimately lifetime prediction, to evaluate the effect of HM stream on the prediction performance.

4.5. Results and analysis

4.5.1. Analysis of smoothing results. Since the vibration signals of the rolling bearings are affected by noise and working conditions, they tend to oscillate considerably and are not favorable for RUL prediction. Therefore, in order to suppress the influence of noise and outliers and reduce the interference

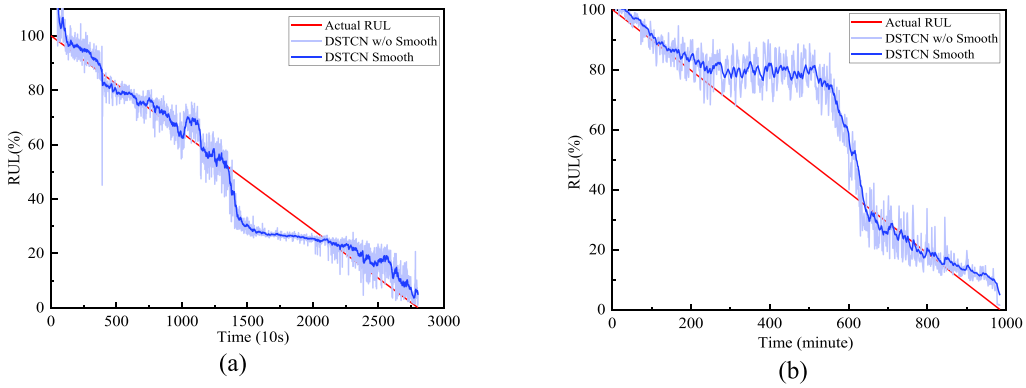


Figure 7. Comparison of DSTCN prediction smoothing filtering. (a) A1–1, (b) B1–1.

Table 4. Performance comparison of different methods on the FEMTO dataset.

		Test bearing													
Methods	Metric	A1–1	A1–2	A1–3	A1–4	A1–5	A1–6	A1–7	A2–1	A2–2	A2–3	A2–4	A2–5	A2–6	A2–7
DSTCN	RMSE	**8.33**	**9.06**	7.24	**13.80**	**10.01**	9.69	**13.62**	**15.30**	**19.07**	**17.59**	24.23	**19.33**	**13.80**	**11.75**
	MAE	**6.16**	**7.29**	6.54	**12.00**	**8.02**	**7.90**	10.97	**12.42**	**15.69**	**13.83**	21.07	**16.14**	**9.61**	**10.46**
	Score	**0.78**	**0.77**	**0.65**	**0.58**	**0.73**	**0.62**	**0.66**	**0.46**	**0.50**	**0.57**	0.28	**0.44**	0.49	**0.43**
TCN	RMSE	10.58	9.77	10.54	17.36	12.98	13.56	14.64	20.53	19.12	21.28	22.45	36.23	18.66	21.94
	MAE	8.54	9.26	8.88	13.13	10.22	10.29	10.13	16.91	15.88	17.22	17.93	29.07	15.51	19.32
	Score	0.63	0.59	0.43	0.49	0.40	0.50	0.64	0.45	0.42	0.40	0.32	0.37	0.44	0.34
TCN-SA [44]	RMSE	15.69	14.30	7.61	14.86	20.38	**8.70**	15.96	23.50	19.75	22.16	**20.54**	23.31	18.16	21.00
	MAE	13.06	11.59	6.45	12.65	16.04	7.91	11.73	19.30	17.08	18.22	**17.12**	19.62	14.90	17.55
	Score	0.59	0.65	0.59	0.51	0.25	0.60	0.63	0.45	0.45	0.40	**0.37**	0.42	0.47	0.37
TCN-RSA [36]	RMSE	13.65	11.75	**7.07**	15.31	14.49	20.59	13.89	24.80	25.84	24.22	24.88	31.85	22.27	18.32
	MAE	11.62	9.08	**5.85**	13.26	10.05	17.24	**9.97**	21.59	22.71	21.22	21.58	26.69	18.67	16.13
	Score	0.66	0.55	0.63	0.51	0.41	0.53	0.60	0.41	0.38	0.39	0.32	0.39	0.37	0.41
CDCT [38]	RMSE	14.43	15.94	11.87	13.99	10.94	11.41	11.42	24.18	23.95	17.78	25.46	25.47	16.7	19.44
	MAE	12.72	14.10	10.66	15.64	8.86	9.69	9.20	20.35	21.23	15.60	21.09	22.20	15.11	16.04
	Score	0.63	0.64	0.44	0.46	0.68	0.55	0.54	0.46	0.17	0.39	0.32	0.34	**0.55**	0.28

with the original data, this paper introduces Kalman filtering for real-time smoothing of DSTCN prediction results. The specific process is as follows:

$$\widehat{\mathrm{RUL}}_t = \widehat{\mathrm{RUL}}_{t-1} + k \cdot \left(\mathrm{RUL}_t - \widehat{\mathrm{RUL}}_{t-1}\right) \tag{31}$$

$$k = \frac{P_{t-1} + Q}{P_{t-1} + Q + R} \tag{32}$$

where $\widehat{\mathrm{RUL}}_t$ is the Kalman filtered value and RUL_t is the unfiltered value. P_{t-1} is the error covariance and $P_0 = 1$. Q is the predicted state covariance, and R is the observed state covariance.

The filtering results are shown in figure 7. In figure 7, the DSTCN prediction results can better reflect the deterioration trend of the bearing during operation, but there are oscillations. The introduction of Kalman filter can make the DSTCN prediction smoother and reduce the influence of oscillation on RUL prediction.

4.5.2. RUL prediction results for the FEMTO dataset. The experimental results are shown in tables 4 and 5. From table 4, DSTCN outperforms the other methods in most of the cases, proving the superiority of the proposed model. The poor performance of TCN and TCN-SA for RUL prediction proves that the original vibration signals as inputs result in the models that cannot easily capture the information of bearing degradation.

The prediction results of TCN-RSA can be seen that the use of Hilbert marginal spectra only as the input would ignore some degradation information of the vibration signal in the time domain. CDCT obtains better RUL prediction results by using both time and frequency domains as inputs. However, the original vibration signals are non-smooth. Fast Fourier Transform itself is defective in dealing with non-smooth signals.

As can be seen from table 5, the mean values of MAE, RMSE and Score of DSTCN are better than the comparison methods on the FEMTO dataset. Compared with TCN and TCN-RSA, the prediction error of DSTCN is much smaller, where MAE decreases by 4.06 and 5.43, respectively.

Table 5. Comparison of the average results of different methods for the FEMTO dataset.

Methods	RMSE	Difference	MAE	Difference	Score	Difference
DSTCN	13.77	—	11.29	—	0.56	—
TCN	17.83	4.06↑	14.44	3.15↑	0.45	0.11↓
TCN-SA [44]	17.56	3.79↑	14.76	3.47↑	0.48	0.08↓
TCN-RSA [36]	19.20	5.43↑	16.11	4.82↑	0.46	0.10↓
CDCT [38]	17.35	3.58↑	15.17	3.88↑	0.46	0.10↓

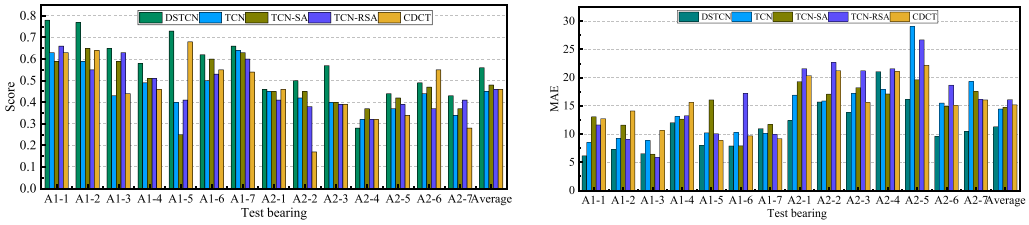


Figure 8. MAE and Score comparing different methods on the FEMTO dataset.

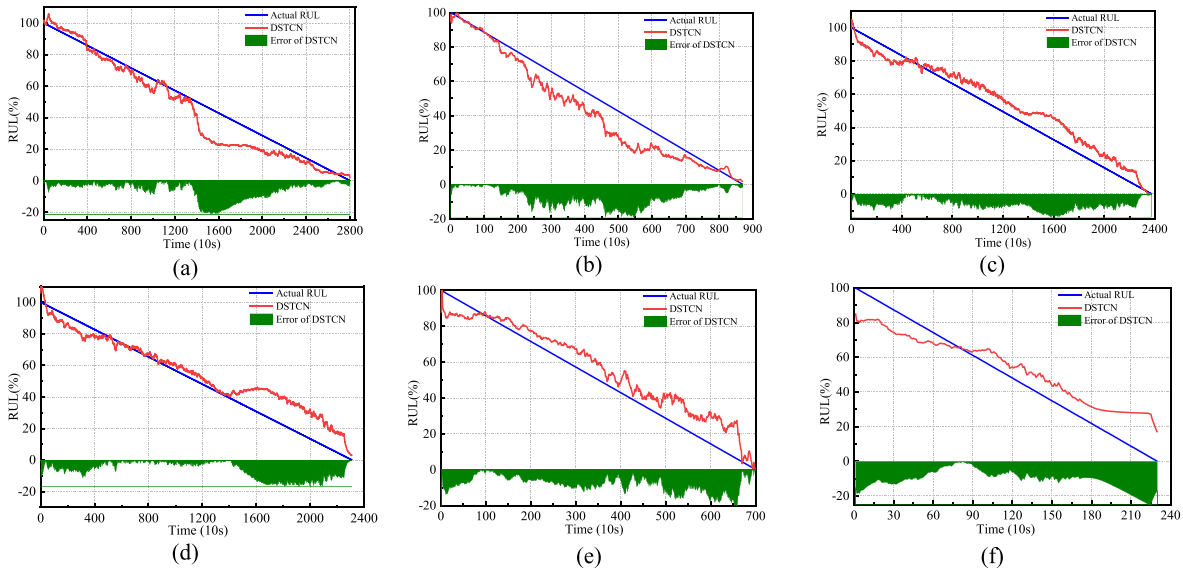


Figure 9. RUL prediction results for the FEMTO dataset: (a) A1–1; (b) A1–2; (c) A1–3; (d) A2–5; (e) A2–6; (f) A2–7.

Compared with TCN-SA and CDCT, the Score of DSTCN is improved by 0.08% and 0.10%.

For a more intuitive comparison, the results of MAE and Score of different methods in table 4 are shown in figure 8, as can be seen, for MAE, except for A1–3 and A2–4, MAE of DSTCN is significantly smaller than that of other comparison methods, which indicates that the predicted RUL curve of the proposed method always fluctuates around the real curve and the error region is smaller. The average value of MAE of DSTCN is also much smaller than its comparison methods. For the Score values, except for A2–4 and A2–6, the Score values of DSTCN are significantly higher than those of the comparison methods, which indicate that the proposed method can effectively capture the late operating condition of the equipment, which is important for early termination of the equipment. In addition, it can be seen that the values of test bearings in working condition 1 are all higher, which is because the speed and applied load in working condition 1 are smaller and the data difference is not too big, and DSTCN is able to capture the test bearing degradation performance more strongly.

Figure 9 shows the results of RUL prediction for the FEMTO dataset. In figure 9, the RUL values of DSTCN always fluctuate around the actual RUL values, which can well reflect the degradation trend of the tested bearings. However, DSTCN prediction results are more accurate in the early stage of the whole life cycle, and the evaluation errors are mainly distributed in the late stage of the life cycle. This is because we assume that the RUL decay process for the test bearings is linear, and this assumption is reasonable during healthy operation of the equipment. However, in the real industry, the occurrence of early equipment failures until complete failure is nonlinear, especially, the large variation of data from equipment operation to the late stage, which leads to DSTCN not following real RUL well.

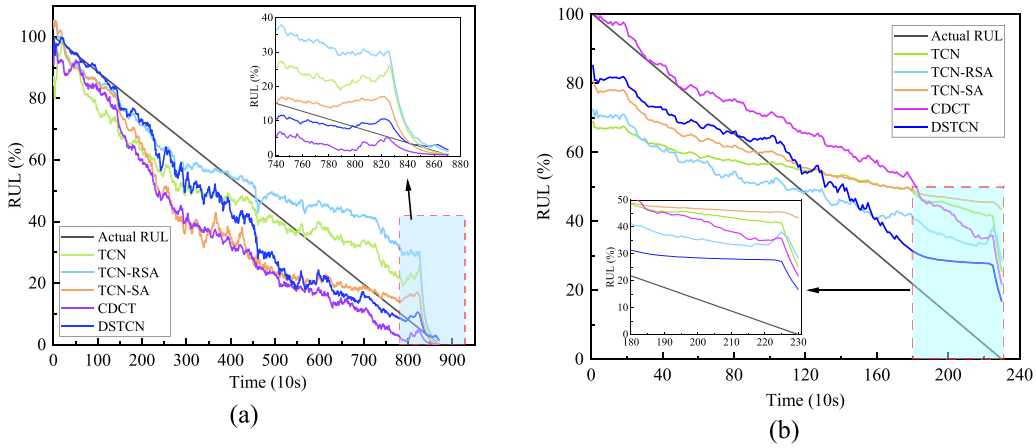


Figure 10. Visualization results of different methods for the FEMTO dataset (a) A1–1; (b) A2–7.

Table 6. Performance comparison of different methods on the XJTU-SY dataset.

		Test bearing									
Methods	Metric	B1–1	B1–2	B1–3	B1–4	B1–5	B2–1	B2–2	B2–3	B2–4	B2–5
DSTCN	RMSE	**12.43**	**11.71**	12.16	**23.87**	**17.49**	**21.27**	**12.3**	**11.42**	**17.72**	**14.13**
	MAE	**8.84**	**9.75**	9.77	**19.31**	**13.86**	**18.78**	**9.57**	**8.92**	**13.94**	**11.26**
	Score	**0.55**	**0.38**	**0.65**	**0.48**	**0.43**	**0.42**	**0.58**	**0.59**	**0.59**	**0.45**
TCN	RMSE	15.07	20.57	18.55	39.26	19.82	36.00	19.99	15.47	29.24	21.78
	MAE	11.81	16.61	16.15	31.60	16.16	28.85	17.45	12.51	22.8	18.46
	Score	0.52	0.34	0.35	0.35	0.39	0.36	0.50	0.51	0.53	0.40
TCN-SA[44]	RMSE	14.58	16.08	**11.46**	36.61	25.66	28.79	12.71	13.98	26.01	17.59
	MAE	10.87	13.35	**8.09**	30.28	19.76	23.73	10.53	11.26	19.63	14.41
	Score	0.47	0.37	0.61	0.37	0.36	0.35	0.55	0.54	0.52	0.32
TCN-RSA[36]	RMSE	22.75	18.16	23.18	32.09	25.42	29.03	13.28	15.73	22.86	15.68
	MAE	17.33	15.7	18.16	27.21	19.89	23.77	11.67	12.88	17.27	12.71
	Score	0.48	0.25	0.48	0.41	0.32	0.36	0.47	0.46	0.52	0.47
CDCT[38]	RMSE	12.63	13.6	16.25	31.37	18.83	35.33	38.63	18.54	43.19	19.83
	MAE	10.14	10.95	12.92	25.83	15.88	28.73	31.43	16.10	35.58	17.05
	Score	0.48	0.38	0.54	0.40	0.41	0.40	0.24	0.55	0.31	0.45

Figure 10 shows the comparison of the prediction results of different methods. From figure 10, the proposed method has higher prediction accuracy and smaller prediction error. It is noteworthy that the comparison methods show significant prediction fluctuations in the late prediction period. From figure 10(a), the proposed method can follow the real RUL better in the early stage of equipment life. In the late stage of the equipment life, although there is a large error, the prediction results are still the best among all the methods. From figure 10(b), the prediction result of DSTCN is the best among all the methods, which further validates the effectiveness of DSTCN for RUL prediction.

4.5.3. RUL prediction results for the XJTU-SY dataset. In this section, the prediction performance of DSTCN and the other four comparative methods were evaluated by three metrics. The results are shown in tables 6 and 7. As can be seen from table 6, except for the RMSE and MAE of test bearing B1–3, all other evaluation metrics of DSTCN are higher than those of the comparison methods, which further proves the superiority of the proposed method. From table 7, the average value of each evaluation index of DSTCN is higher than that of other comparison methods. Among them, MAE of DSTCN is 6.84, 3.79, 5.25 and 8.06 lower than that of TCN, TCN-SA, TCN-RSA and CDCT, respectively. The above results indicate that TCN has a simple structure and does not use optimization techniques, resulting in capturing the bearing degradation performance bad. TCN-SA and TCN-RSA have RUL prediction capability. However, TCN-SA uses time domain information as input and TCN-RSA uses Hilbert marginal spectrum as input, which are single-input models, resulting in limited input information and poor RUL prediction. CDCT uses time domain and frequency domain signals after fast Fourier transform (FFT) as input, but FFT itself has some defects in handling non-stationary signals, resulting in poor prediction of

Table 7. Comparison of the average results of different methods for the XJTU-SY dataset.

Methods	RMSE	Difference	MAE	Difference	Score	Difference
DSTCN	15.45	—	12.4	—	0.51	—
TCN	23.57	8.12↑	19.24	6.84↑	0.42	0.09↓
TCN-SA [44]	20.34	4.89↑	16.19	3.79↑	0.45	0.06↓
TCN-RSA [36]	21.81	6.36↑	17.65	5.25↑	0.42	0.09↓
CDCT [38]	24.82	9.37↑	20.46	8.06↑	0.41	0.10↓

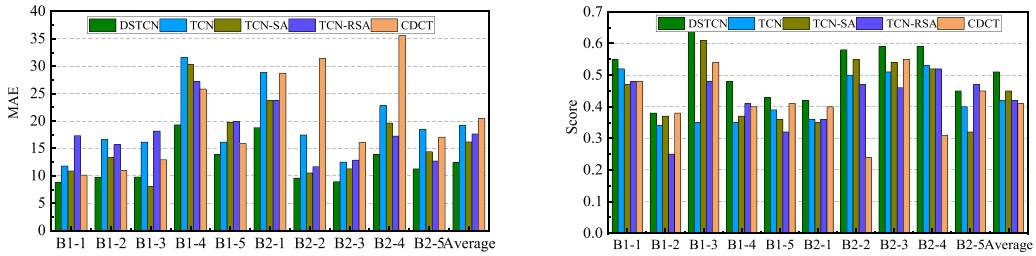


Figure 11. MAE and Score comparing different methods on the XJTU-SY dataset.

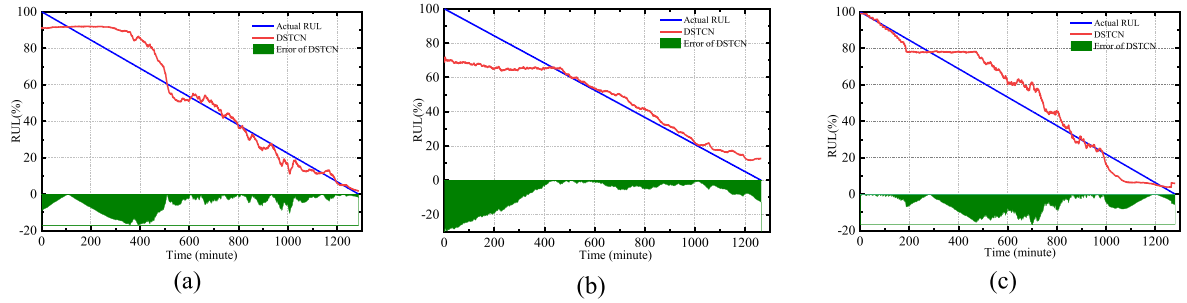


Figure 12. RUL prediction results for the XJTU-SY dataset: (a) B1–2; (b) B1–3; (c) B2–3.

CDCT. For a more intuitive comparison, the MAE and Score values of the different methods in table 7 are visualized, and the results are shown in figure 11. The average value of MAE of DSTCN is also much smaller than that of its comparison methods. DSTCN has a higher score value than the comparison methods, which indicates that the proposed method can effectively capture the late operating status of the equipment, which is important for early termination of the equipment.

Figure 12 shows the RUL prediction visualization results for the XJTU-SY dataset. In figure 12, DSTCN can capture the degradation trend of different test bearings. RUL curves of DSTCN have large errors with the true RUL curves. This is because we assume that the RUL degradation process of the test bearings is linear, which is reasonable during the healthy operation period of the equipment. However, in the real industry, the early failure of the equipment until complete failure is nonlinear. Especially, the large variation of the data from the equipment operation to the later period, which leads to DSTCN not following the real RUL well.

Figure 13 shows the comparison of the prediction results of different methods, and it can be seen that the proposed method can follow the real RUL degradation trend better, with higher prediction accuracy and lower prediction error. In particularly, it is worth noting that the fluctuation between the prediction

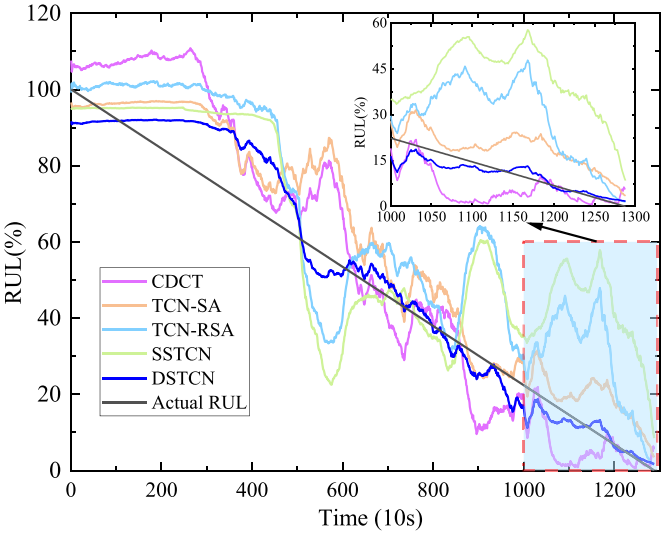


Figure 13. Test bearing B1-2 RUL prediction results of different methods.

results of DSTCN and the real RUL curve at the late stage of equipment life is small and the error is low. On the contrary, the comparative methods have larger fluctuations in the prediction

Table 8. Detailed description of ablation and classical methods.

Methods	Preprocessing module		Feature extraction module
Ablation methods	Original vibration signal	HMs signal	Hybrid attention
STCN	√		√
HMSTCN		√	√
DSTCN-WHAT	√	√	
DSTCN	√	√	√
Classical methods			
LSTM	√	√	√
BiGRU	√	√	√

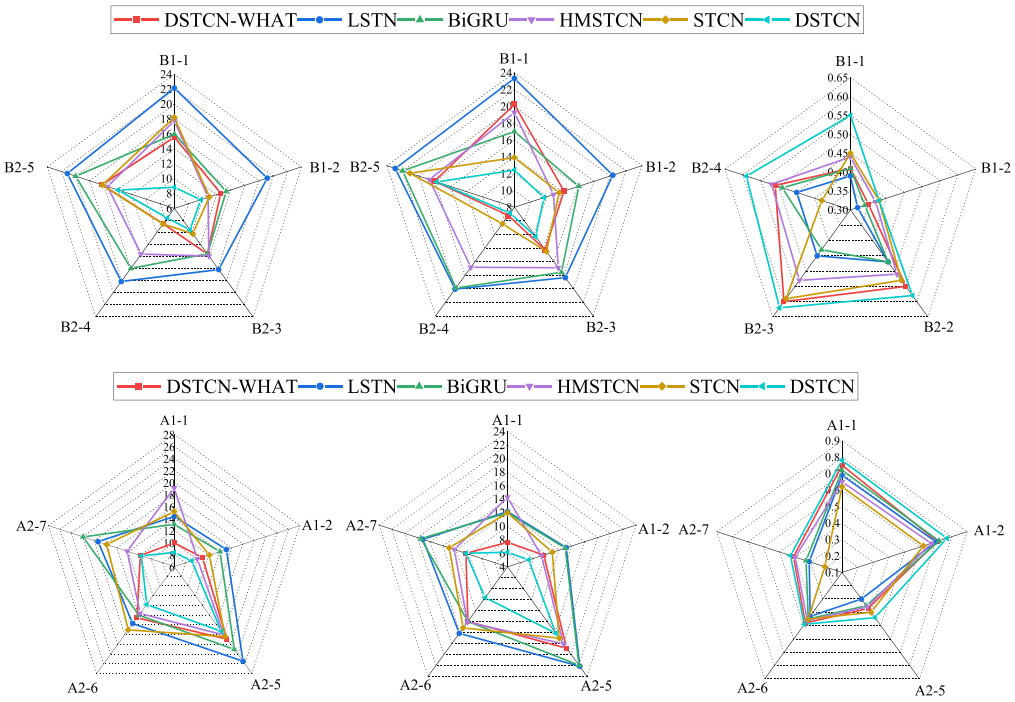


Figure 14. Radar plot of MAE, RMSE and Score.

results at the late stage of prediction, and the prediction effect is poorer.

4.5.4. Ablation experiments. To verify the effect of each module in DSTCN, we selected five types of bearing data from the above two datasets for ablation experiments. Specifically, the factors affecting the performance of DSTCN include: (1) single-stream raw vibration signals (STCN); (2) single-stream HM (HMSTCN); (3) TCN without hybrid attention (DSTCN-WHAT). In addition, to further verify the advancement of the proposed method, we select LSTN and BiGRU for comparison. The specific details are shown in table 8. To reduce the randomness caused by parameter initialization, the experimental results obtained are the average of 10 experimental results. Figure 14 shows the performance of different ablation methods and classical methods predicted on the bearing dataset. By observing the performance of different methods on the same bearing dataset, we can find that MAE and RMSE of STCN are higher than those of HMSTCN.

This indicates that the HMs transformation makes the bearing degradation information more obvious, which helps the model to obtain the degradation features more easily. In addition, when the model uses dual-stream inputs without hybrid attention, the Score of DSTCN-WHAT are lower than those of DSTCN. This indicates that hybrid attention can help the model to focus on the information between different channels and increase the relevance of long-distance feature information. For the classical methods, the prediction performance of BiGRU is better than that of LSTM. This is because that BiGRU can capture the degeneracy information better using bidirectional recurrent structure.

Figure 15 shows the running time of the ablation and classical methods for training a batch. We can see that DSTCN has the longest training time among the three ablation methods. This is because STCN and HMSTCN are single-stream models and the number of parameters in the model is small. However, the training time of DSTCN is much lower than that of LSTN and BiGRU. This indicates that DSTCN is advantageous in terms of timeliness.

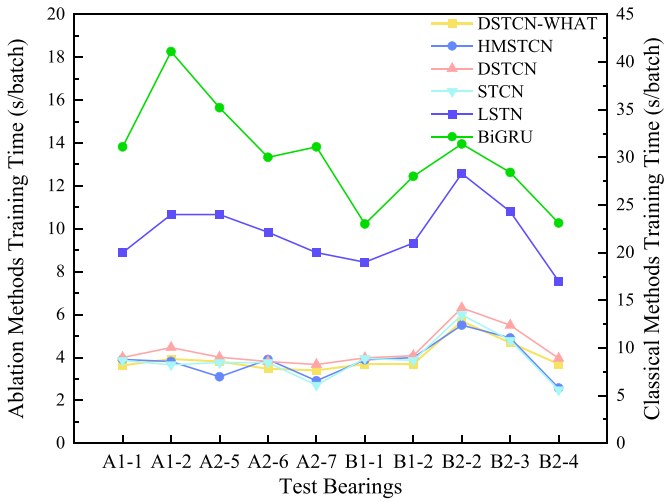


Figure 15. Training time for different methods.

5. Conclusion

In this paper, a DSTCN is proposed for RUL prediction of rolling bearings. Firstly, DSTCN takes HM and vibration signals as network inputs. Secondly, a HATCB is designed by combining the advantages of dilated causal convolution and attention mechanisms. Degraded features of different input signals are captured by multiple stacked HATCB. Thirdly, feature recalibration and assignment are performed by using one-dimensional attentional fusion module before the fully connected layer. Finally, the effectiveness of the proposed method is validated with two life-cycle datasets, and the results show that the proposed method has better accuracy than other methods in RUL prediction.

In future, we will consider multiple failure modes and complex operating conditions to research RUL prediction methods.

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.

Acknowledgment

This work was financially supported by the National Natural Science Foundation of China (No. 61763029), the College Industrial Support Project of Gansu Province (2023CYZC-24), the Science and Technology Project of Gansu Province (24JRRA172)

Conflict of 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.

ORCID iDs

Yazhou Zhang https://orcid.org/0000-0002-2876-0330
Xiaoqiang Zhao https://orcid.org/0000-0001-5687-942X
Zhenrui Peng https://orcid.org/0000-0003-1140-4900

References

[1] Liang H, Cao J and Zhao X 2022 Multi-scale dynamic adaptive residual network for fault diagnosis *Measurement* **188** 110397
[2] Chang Y, Chen J, Liu Y, Xu E and He S 2022 Temporal convolution-based sorting feature repeat-explore network combining with multi-band information for remaining useful life estimation of equipment *Knowl.-Based Syst.* **249** 108958
[3] Mou M and Zhao X 2022 Gated broad learning system based on deep cascaded for soft sensor modeling of industrial process *IEEE Trans. Instrum. Meas.* **71** 1–11
[4] Cao Y, Jia M, Ding Y, Zhao X, Ding P and Gu L 2023 Complex domain extension network with multi-channels information fusion for remaining useful life prediction of rotating machinery *Mech. Syst. Signal Process.* **192** 110190
[5] Li T, Zhao Z, Sun C, Yan R and Chen X 2021 Hierarchical attention graph convolutional network to fuse multi-sensor signals for remaining useful life prediction *Reliab. Eng. Syst. Saf.* **215** 107878
[6] Chen D, Qin Y, Wang Y and Zhou J 2021 Health indicator construction by quadratic function-based deep convolutional auto-encoder and its application into bearing RUL prediction *ISA Trans.* **114** 44–56
[7] Zhao Z, Liang B, Wang X and Lu W 2017 Remaining useful life prediction of aircraft engine based on degradation pattern learning *Reliab. Eng. Syst. Saf.* **164** 74–83
[8] Arunan A, Qin Y, Li X and Yuen C 2024 A change point detection integrated remaining useful life estimation model under variable operating conditions *Control Eng. Pract.* **144** 105840
[9] Chang Y, Chen J, Chen Q, Liu S and Zhou Z 2022 CFs-focused intelligent diagnosis scheme via alternative kernels networks with soft squeeze-and-excitation attention for fast-precise fault detection under slow & sharp speed variations *Knowl.-Based Syst.* **239** 108026
[10] He M, Zhou Y, Li Y, Wu G and Tang G 2020 Long short-term memory network with multi-resolution singular value decomposition for prediction of bearing performance degradation *Measurement* **156** 107582
[11] Zhang X, Guo Y, Shangguan H, Li R, Wu X and Wang A 2023 Predicting remaining useful life of a machine based on embedded attention parallel networks *Mech. Syst. Signal Process.* **192** 110221
[12] Chen Y, Zhang D and Zhang W-A 2022 MSWR-LRCN: a new deep learning approach to remaining useful life estimation of bearings *Control Eng. Pract.* **118** 104969
[13] Li W, Shang Z, Gao M, Qian S and Feng Z 2022 Remaining useful life prediction based on transfer multi-stage shrinkage attention temporal convolutional network under variable working conditions *Reliab. Eng. Syst. Saf.* **226** 108722
[14] Chang Y, Li F, Chen J, Liu Y and Li Z 2022 Efficient temporal flow transformer accompanied with multi-head probsparse self-attention mechanism for remaining useful life prognostics *Reliab. Eng. Syst. Saf.* **226** 108701
[15] Yang Q, Tang B, Yang S and Shen Y 2023 An integrated network architecture for data repair and degradation trend prediction *Mech. Syst. Signal Process.* **200** 110610

[16] Deng Y, Di Bucchianico A and Pechenizkiy M 2020 Controlling the accuracy and uncertainty trade-off in RUL prediction with a surrogate Wiener propagation model *Reliab. Eng. Syst. Saf.* **196** 106727
[17] Chen D, Qin Y, Qian Q, Wang Y and Liu F 2023 Transfer life prediction of gears by cross-domain health indicator construction and multi-hierarchical long-term memory augmented network *Reliab. Eng. Syst. Saf.* **230** 108916
[18] Zang Y, Shangguan W, Cai B, Wang H and Pecht M G 2021 Hybrid remaining useful life prediction method. A case study on railway D-cables *Reliab. Eng. Syst. Saf.* **213** 107746
[19] Jouin M, Gouriveau R, Hissel D, Péra M-C and Zerhouni N 2016 Particle filter-based prognostics: review, discussion and perspectives *Mech. Syst. Signal Process.* **72** 2–31
[20] Yu W, Tu W, Kim I Y and Mechefske C 2021 A nonlinear-drift-driven Wiener process model for remaining useful life estimation considering three sources of variability *Reliab. Eng. Syst. Saf.* **212** 107631
[21] Jouin M, Gouriveau R, Hissel D, Péra M-C and Zerhouni N 2016 Degradations analysis and aging modeling for health assessment and prognostics of PEMFC *Reliab. Eng. Syst. Saf.* **148** 78–95
[22] Khelif R, Chebel-Morello B, Malinowski S, Laajili E, Fnaiech F and Zerhouni N 2016 Direct remaining useful life estimation based on support vector regression *IEEE Trans. Ind. Electron.* **64** 2276–85
[23] Gebraeel N, Lawley M, Liu R and Parmeshwaran V 2004 Residual life predictions from vibration-based degradation signals: a neural network approach *IEEE Trans. Ind. Electron.* **51** 694–700
[24] Hinton G E, Osindero S and Teh Y-W 2006 A fast learning algorithm for deep belief nets *Neural Comput.* **18** 1527–54
[25] Wei Y, Wu D and Terpenny J 2024 Remaining useful life prediction using graph convolutional attention networks with temporal convolution-aware nested residual connections *Reliab. Eng. Syst. Saf.* **242** 109776
[26] Xia P *et al* 2023 Adaptive feature utilization with separate gating mechanism and global temporal convolutional network for remaining useful life prediction *IEEE Sens. J.* **23** 21408–20
[27] Tian H, Yang L and Ju B 2023 Spatial correlation and temporal attention-based LSTM for remaining useful life prediction of turbofan engine *Measurement* **214** 112816
[28] Shi J, Zhong J, Zhang Y, Xiao B, Xiao L and Zheng Y 2024 A dual attention LSTM lightweight model based on exponential smoothing for remaining useful life prediction *Reliab. Eng. Syst. Saf.* **243** 109821
[29] Qin Y, Chen D, Xiang S and Zhu C 2020 Gated dual attention unit neural networks for remaining useful life prediction of rolling bearings *IEEE Trans. Ind. Inform.* **17** 6438–47
[30] Qiu H, Niu Y, Shang J, Gao L and Xu D 2023 A piecewise method for bearing remaining useful life estimation using temporal convolutional networks *J. Manuf. Syst.* **68** 227–41
[31] Zhu G, Zhu Z, Xiang L, Hu A and Xu Y 2023 Prediction of bearing remaining useful life based on DACN-ConvLSTM model *Measurement* **211** 112600
[32] Sun J, Zhang X and Wang J 2023 Lightweight bidirectional long short-term memory based on automated model pruning with application to bearing remaining useful life prediction *Eng. Appl. Artif. Intell.* **118** 105662
[33] Zhang L, Wang B, Yuan X and Liang P 2022 Remaining useful life prediction via improved CNN, GRU and residual attention mechanism with soft thresholding *IEEE Sens. J.* **22** 15178–90
[34] Shen Y, Tang B, Li B, Tan Q and Wu Y 2022 Remaining useful life prediction of rolling bearing based on multi-head attention embedded Bi-LSTM network *Measurement* **202** 111803
[35] Wang W *et al* 2023 A novel competitive temporal convolutional network for remaining useful life prediction of rolling bearings *IEEE Trans. Instrum. Meas.* **72** 3523612
[36] Cao Y, Ding Y, Jia M and Tian R 2021 A novel temporal convolutional network with residual self-attention mechanism for remaining useful life prediction of rolling bearings *Reliab. Eng. Syst. Saf.* **215** 107813
[37] Shi C, Luo B, He S, Li K, Liu H and Li B 2019 Tool wear prediction via multidimensional stacked sparse autoencoders with feature fusion *IEEE Trans. Ind. Inform.* **16** 5150–9
[38] Jiang L, Zhang T, Lei W, Zhuang K and Li Y 2023 A new convolutional dual-channel Transformer network with time window concatenation for remaining useful life prediction of rolling bearings *Adv. Eng. Inform.* **56** 101966
[39] Xu D, Qiu H, Gao L, Yang Z and Wang D 2022 A novel dual-stream self-attention neural network for remaining useful life estimation of mechanical systems *Reliab. Eng. Syst. Saf.* **222** 108444
[40] Fu K, Qu J, Chai Y and Zou T 2015 Hilbert marginal spectrum analysis for automatic seizure detection in EEG signals *Biomed. Signal Process. Control* **18** 179–85
[41] Evensen G 2003 The ensemble Kalman filter: theoretical formulation and practical implementation *Ocean Dyn.* **53** 343–67
[42] Li X, Wang W, Hu X and Yang J 2019 Selective kernel networks *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 510–9
[43] Nectoux P *et al* 2012 PRONOSTIA: an experimental platform for bearings accelerated degradation tests *IEEE Int. Conf. on Prognostics and Health Management, PHM'12* pp 1–8
[44] Wang Y, Deng L, Zheng L and Gao R X 2021 Temporal convolutional network with soft thresholding and attention mechanism for machinery prognostics *J. Manuf. Syst.* **60** 512–26

控制与决策
Control and Decision
ISSN 1001-0920,CN 21-1124/TP

《控制与决策》网络首发论文

题目：基于多传感器数据融合的 SA-DACNN 齿轮箱故障诊断方法
作者：张亚洲，赵小强，惠永永，陈鹏
DOI：10.13195/j.kzyjc.2023.1367
收稿日期：2023-09-25
网络首发日期：2024-02-04
引用格式：张亚洲，赵小强，惠永永，陈鹏．基于多传感器数据融合的 SA-DACNN 齿轮箱故障诊断方法[J/OL]．控制与决策. https://doi.org/10.13195/j.kzyjc.2023.1367

基于多传感器数据融合的 SA-DACNN 齿轮箱故障诊断方法

张亚洲[1,2]，赵小强[1,2†]，惠永永[1,2]，陈 鹏[3]

(1. 兰州理工大学电气工程与信息工程学院，兰州 730050；2. 甘肃省工业过程先进控制重点实验室, 兰州 730050；3. 兰州石化职业技术大学电子电气工程学院, 兰州 730050)

摘 要: 针对单一传感器数据易受自身品质和环境的影响导致难以监控齿轮箱整体运行状况的问题, 本文提出一种基于多传感器数据融合的 SA-DACNN（Self Attention - Dynamic Adaptive Convolutional Neural Network）齿轮箱故障诊断方法. 首先, 该方法将采集到的不同位置的传感器信号作为多通道信号, 并将多通道信号同时作为网络输入; 然后, 设计了一种多通道特征融合模块, 该模块通过自适应地加权不同通道的信息, 确保不同通道的重要信息能够有效地融合, 解决了特征级多通道数据融合问题; 最后, 在全连接层之前, 使用带残差连接的自注意力模块, 帮助网络自动学习全局信息, 增强对原始振动信号的特征学习能力. 在两个齿轮箱数据集中进行实验, 结果表明, 本文所提方法具有较高的故障诊断准确率, 可以满足多传感器数据融合故障诊断的任务.

关键词: 故障诊断；多传感器数据；数据融合；齿轮箱；卷积神经网络；自注意力

中图分类号: TH133.3;TP206.3 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2023.1367

引用格式: 张亚洲，赵小强，惠永永，等. 基于多传感器数据融合的 SA-DACNN 齿轮箱故障诊断方法 [J]. 控制与决策.

SA-DACNN gearbox fault diagnosis method based on multi-sensor data fusion

Zhang Ya-zhou[1,2], *Zhao Xiao-qiang*[1,2†], *Hui Yong-yong*[1,2], *Chen Peng*[3]

(1.College of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China; 2.Key Laboratory of Advanced Control of Industrial Processes in Gansu Province, Lanzhou 730050, China; 3.College of Electrical and Electronic Engineering, Lanzhou Petrochemical University of Technology, Lanzhou 730050, China)

Abstract: To address the problem that single sensor data are easily affected by their own quality and environment, which makes it difficult to monitor the overall operating condition of gearboxes, a SA-DACNN (Self Attention - Dynamic Adaptive Convolutional Neural Network) gearbox fault diagnosis method based on multi-sensor data fusion is proposed in this paper. Firstly, the method treats the collected sensor signals from different locations as multi-channel signals and uses the multi-channel signals as network inputs simultaneously. Secondly, a multi-channel feature fusion module is designed, which solves the feature-level multi-channel data fusion problem by adaptively weighting the information of different channels to ensure that the important information of different channels can be effectively fused. Finally, before the fully connected layer, a self-attentive module with residual connections is used to help the network automatically learn global information and enhance the feature learning ability of the original vibration signals. Experiments are conducted on two gearbox datasets, and the results show that the proposed method has a high fault diagnosis accuracy and can meet the task of multi-sensor data fusion fault diagnosis.

Keywords: Fault diagnosis; Multi-sensor data; Data fusion; Gearbox; Convolutional neural network; Self attention

0 引 言

齿轮箱作为机械传动系统的核心部件, 其可靠性对于设备的安全运行至关重要. 由于重要部位的齿轮发生故障, 可能会阻碍机械系统进行正常生产并导致经济损失. 因此, 如何对齿轮箱进行有效故障检测和诊断, 及时发现设备的故障, 防止出现计划外停机并减少昂贵的维修费用具有重大的研究意义[1-2].

随着传感器技术的发展, 数据采集变得更加方便, 因此, 基于数据驱动的深度学习方法在机械故障

收稿日期: 2023-09-25；**录用日期:** 2024-01-16.

基金项目: 国家自然科学基金项目(62263021)；甘肃省青年科技基金计划（22JR5RA808）.

†通讯作者. E-mail: xqzhao@lut.edu.cn.

诊断中得到快速发展[3-5]. 然而, 单一传感器获得的信息有限且容易受到外界因素的干扰, 难以监控齿轮箱的整体状况. 为了全面了解齿轮箱的状况, 提升故障诊断精度对齿轮箱布置多个传感器采集信息, 并通过多源信息融合技术将多个传感器信息进行融合, 最终获得互补信息已成为机械设备故障诊断发展的研究热点[6].

多传感器信息融合包含三个策略级融合: 数据级融合、特征级融合以及决策级融合[7]. 数据级融合是指将多个传感器采集的原始数据在网络输入层之前直接进行融合. 例如:Jing 等人[8] 将多个传感器采集的原始数据直接融合, 提出了一种基于深度卷积神经网络的行星齿轮箱故障诊断方法. Azamfar 等人[9] 将多个电流传感器采集的原始信号进行融合, 然后使用二维卷积神经网络进行齿轮箱故障诊断. 数据级融合可以充分利用原始数据克服特征信息丢失, 但是此类方法需要人为选取原始数据, 且缺乏可解释性. 决策级融合是指将每个传感器的特征信息在网络的输出层进行融合. 例如: 杨等人[10] 将多个传感器数据输入到一维卷积神经进行特征提取, 然后在全连接层将每个传感器的特征信息进行融合, 最终实现航空发动机轴承故障诊断. Gong 等人[11] 提出了基于改进卷积神经网络和多传感器数据融合的快速异常诊断方法, 该方法以多传感器数据为输入, 在全连接层进行特征融合. 决策级融合拥有较好的可解释性和实用性, 但是此类方法进行特征融合时信息丢失严重, 故障诊断效果不理想. 特征级融合是指使用特征选择方法在特征提取过程中进行信息融合. 例如:Xu 等人[12] 提出了一种基于加权软投票规则的多注意力模块多尺度信息融合的机械维修智能故障诊断方法, 该方法充分考虑了不同尺度的特征对网络决策层的贡献, 解决了多尺度网络在复杂环境下故障诊断低效的问题. Zhang 等人[13] 提出了一种计算资源较少的多传感器振动信号故障诊断方法, 该方法在特征层进行多层特征融合, 并使用多层池化分类器取代传统卷积网络的全连接层, 有效地减少了网络参数量和过拟合风险. 特征级融合可以将原始高维数据进行压缩降维, 但是此类方法需要领域知识.

综上可知, 基于多传感器信息融合的故障诊断方法取得了令人鼓舞的成果, 但是仍然存在一些问题, 例如: 数据级融合信息损失小, 但该融合方法研究有限, 不同类型的原始数据进行融合缺乏理论解释; 决策级融合可解释性好, 但是特征信息损失巨大, 诊断效果不佳; 特征级融合尽管依赖领域知识, 但是特征信息损失较小, 易于实现, 所以本文采用多传感器特征级融合策略. 此外, 考虑到深度学习具有较强大的特征提取能力, 不需要手动提取特征. 因此, 本文提出了一种基于多传感器特征级数据融合的 SA-DACNN 故障诊断方法, 该方法首先设计了一种高效的多通道特征融合模块, 解决了特征层面的多通道数据融合问题, 然后为了更好地学习特征信息和标签之间的映射关系, 使用带残差连接的自注意力模块进行全局信息建模, 帮助网络捕获不同通道特征信息对输出结果的贡献度. 与其它方法相比, 本文方法可以动态自适应地融合不同传感器的特征信息, 对专家知识和人力的依赖性较小.

1 一维卷积神经网络（1DCNN）

CNN 是一种前馈深度学习模型, 被广泛用于计算机视觉和模式识别任务中[14-15]. 近年来, 卷积神经网络也被成功用于机械故障诊断领域 [16-18]. 在工业现场, 传感器采集的信号通常为一维时域信号, 因此本文选用一维卷积神经网络作为基本网络. 一维卷积神经网络由卷积层、池化层、激活层和全连接层组成. 卷积层能够对上一层输出进行点积操作, 提取局部区域的特征[19], 其公式为:

$$Z_{l'\times w'}^{j+1}=f\left(W_{l\times w}^{j}\otimes X_{s\times w}^{j}+b^{j}\right) \tag{1}$$

式中, $W_{l\times w}^{j}$ 表示第 j 层卷积核的权重, l 和 w 分别表示卷积核的长度和宽度, $X_{s\times w}^{j}$ 表示上一层输入, s 表示上一层输入的长度, b^{j} 表示第 j 层卷积核的偏差, $\otimes$ 表示点积操作, $Z_{l'\times w'}^{j+1}$ 表示卷积层输出.

池化层能够进一步压缩卷积层输出特征图的维度，降低参数量[20]. 池化层不会改变特征图的宽度, 其公式为:

$$\begin{gathered}P_{l}^{j+1}(i)=\max\left\{a_{w}^{j}(t)\right\}\\ t\in[(i-1)\bullet w+1,i\bullet w]\end{gathered} \tag{2}$$

式中, w 表示池化区域的宽度, $a_{w}^{j}(t)$ 表示第 t 层的第 $P_{l}^{j+1}(i)$ 个神经元, 表示池化层输出值.

全连接层能够将卷积层和池化层输出的特征进行非线性组合, 得到最终的特征向量. 在分类任务中, 使用 Softmax 做分类器, 其公式为:

$$Y_i=\begin{bmatrix}P(y_i=1|x_i)\\ P(y_i=2|x_i)\\ \vdots\\ P(y_i=n|x_i)\end{bmatrix}^{\mathrm{T}}=\frac{1}{\sum\limits_{l=1}^{n}e^{x_i^{\mathrm{T}}\bullet w_l}}\begin{bmatrix}e^{x_i^{\mathrm{T}}\bullet w_1}\\ e^{x_i^{\mathrm{T}}\bullet w_2}\\ \vdots\\ e^{x_i^{\mathrm{T}}\bullet w_n}\end{bmatrix}^{\mathrm{T}} \tag{3}$$

式中, Y_i 表示第 i 个样本 Softmax 归一化输出结果, Y_i

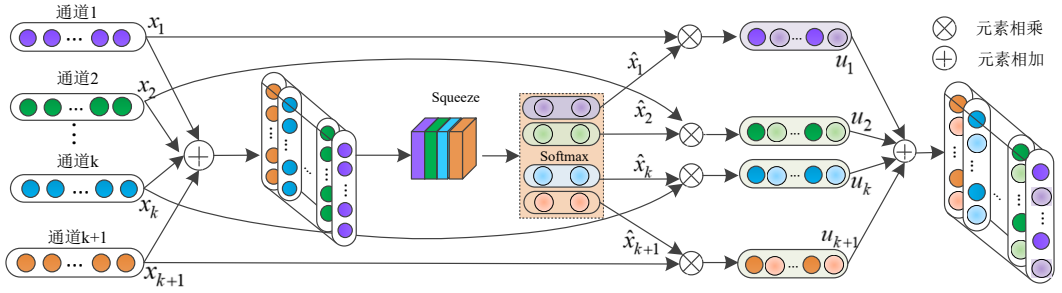


图 1 多通道特征融合模块

输出结果在 $[0,1]$ 之间, $1\Big/\sum_{l=1}^{n} e^{x_i^{\mathrm{T}}\bullet w_l}$ 表示对输出概率 $\left[\, e^{x_1^{\mathrm{T}}\bullet w_1},\ e^{x_1^{\mathrm{T}}\bullet w_2},\ \cdots\ e^{x_1^{\mathrm{T}}\bullet w_n} \,\right]$ 的归一化结果.

2 基于 SA-DACNN 的多传感器数据融合故障诊断方法

2.1 多通道特征融合模块

齿轮箱通常被安装多个传感器监测其运行状态, 单一传感器只能采集固定位置的机械故障信息, 无法捕获齿轮箱的整体运行情况. 为了充分利用不同方向、不同位置的传感器采集的数据, 本文设计了一个多通道特征融合模块, 如图 1 所示, 该多通道特征融合模块通过自适应地加权不同通道的信息, 确保不同通道的重要信息能够有效地融合.

多通道特征融合模块可分为三个步骤, 第一步骤中, 不同通道的特征信息 x_1、$x_2\cdots x_k$、x_{k+1} 进行特征融合 (元素相加) 获得新的特征信息图, 其公式为:

$$U = x_1 + x_2 + ... + x_k + x_{k+1} \tag{4}$$

式中, x_1、$x_2\cdots x_k$ 和 x_{k+1} 表示不同通道的输入特征, 1、2、$\cdots k$ 和 $k+1$ 分别表示第 1 通道、第 2 通道、第 k 通道和第 $k+1$ 通道, U 表示特征融合后的值.

第二步骤中, 首先通过全局平均池化将特征信息展平为一维特征, 实现特征降维, 即 U_{gap}, 然后通过 Softmax 函数加权不同通道的特征信息, 其公式为:

$$\hat{x}_1 = e^{U_{gap}^1} \Big/ \sum_{j=1}^{k+1} e^{U_{gap}^j} \tag{5}$$

$$\hat{x}_2 = e^{U_{gap}^2} \Big/ \sum_{j=1}^{k+1} e^{U_{gap}^j} \tag{6}$$

$$\hat{x}_k = e^{U_{gap}^k} \Big/ \sum_{j=1}^{k+1} e^{U_{gap}^j} \tag{7}$$

$$\hat{x}_{k+1} = e^{U_{gap}^{k+1}} \Big/ \sum_{j=1}^{k+1} e^{U_{gap}^j} \tag{8}$$

$$\hat{x}_1 + \hat{x}_2 + , ..., + \hat{x}_k + \hat{x}_{k+1} = 1 \tag{9}$$

式中, $\hat{x}_1$、$\hat{x}_2\cdots\hat{x}_k$ 和 $\hat{x}_{k+1}$ 表示不同通道特征经过 Softmax 函数归一化后的输出值, $e^{U_{gap}^1}$、$e^{U_{gap}^2}\cdots e^{U_{gap}^k}$ 和 $e^{U_{gap}^{k+1}}$ 分别表示第 1、$2\cdots k$ 和 $k+1$ 通道的转变为 0~1 之间的概率值, $\sum_{j=1}^{k+1} e^{U_{gap}^j}$ 表示对所有通道的概率值求和.

第三步骤中, 将 Softmax 函数输出值和输入特征图相乘, 生成新的加权特征图, 并将新的特征图进行融合, 其公式为:

$$u_i = x_i \bullet \hat{x}_i \tag{10}$$

$$\vec{U} = \sum_{i=1}^{k+1} u_i \tag{11}$$

式中, u_i 表示第 i 通道生成的加权特征图, $\vec{U}$ 表示多通道特征融合模块的输出.

2.2 自注意力模块

目前, 注意力机制被广泛使用于自然语言处理领域, 注意力机制可以帮助网络更加关注对输出有贡献的特征信息[21,22]. 为了使网络更好地学习特征信息和标签之间的映射关系, 本文在全连接输出层前使用带残差连接的自注意力模块 (Residual Self attention module, RSAM), RSAM 能够自动学习特征之间的相关性, 并根据相关性加权处理输入特征, 使其更加关注有用信息, 同时抑制无关信息, 从而提高网络的故障诊断准确率, RSAM 如图 2 所示.

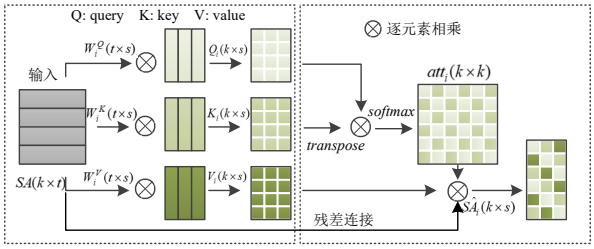


图 2 残差自注意力模块

在图 2 中, RSAM 处理过程可以分为三个步骤:

Step1: 输入特征图 X 经过线性变化分别得到查

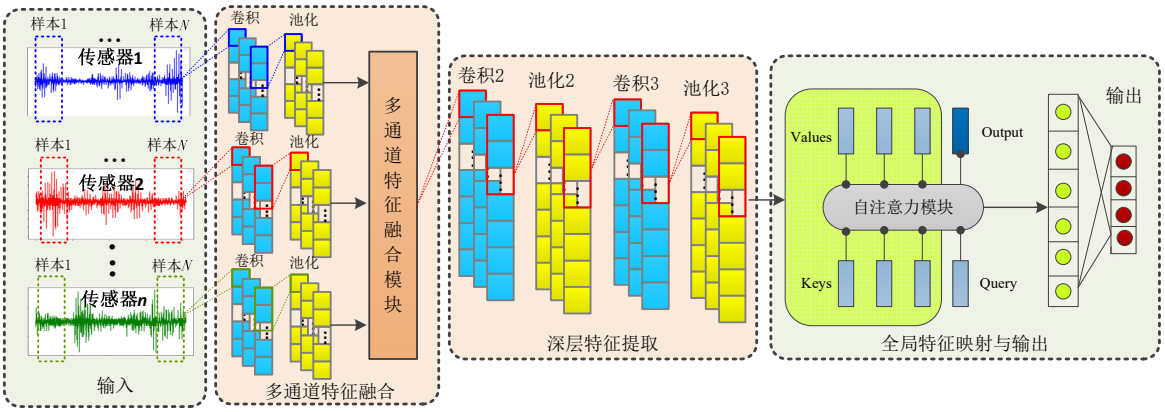


图 3 SA-DACNN 诊断框架

询矩阵 Q_i、键矩阵 K_i 和值矩阵 V_i, 其公式为:

$$Q_i = X \bullet W_i^Q \tag{12}$$

$$K_i = X \bullet W_i^K \tag{13}$$

$$V_i = X \bullet W_i^V \tag{14}$$

式中, W_i^Q、W_i^K、W_i^V 为线性变化的参数矩阵, 参数矩阵通过训练网络不断学习.

Step2: 键矩阵 K_i 进行转置得到 K_i^T, 将 K_i^T 和查询矩阵 Q_i 进行点乘运算得到两个向量之间的相关性, 通过 Softmax 函数计算残差自注意力模块的权重矩阵, 其数学描述为:

$$att_i = softmax(Q_i K_i^T \Big/ \sqrt{d}) \tag{15}$$

式中, d 表示参数矩阵的维数.

Step3: 值矩阵 V_i 和 att_i 进行矩阵乘法运算, 并与输入特征进行残差连接得到自注意力模块的输出, 其数学描述为:

$$S\hat{A}_i = V_i \bullet att_i \tag{16}$$

$$RS\widehat{A}_i = S\widehat{A}_i + SA \tag{17}$$

式中, V_i 表示值矩阵, att_i 表示残差自注意模块的权重矩阵, $S\hat{A}_i$ 表示自注意力输出, SA 表示残差自注意模块的输入.

2.3 多传感器数据融合诊断流程

单一传感器只能获取固定位置的故障信息且容易受到传感器自身性能的影响, 难以体现整个齿轮箱的运行状态. 多源传感器数据融合可以获得更多的故障信息, 为设备运行维护人员提高更加可靠的维修决策. 因此, 本文提出了一种基于多传感器数据融合的 SA-DACNN 故障诊断方法, 该方法无需对原始数据做任何手工处理, 只需将传感器采集的振动信号输入 SA-DACNN 网络, 诊断结果自动输出.

SA-DACNN 诊断网络主要由输入层、特征提取层、多通道特征融合模块、带残差连接的自注意力模块和输出层组成, 网络结构如图 3 所示. 该网络将传统 1DCNN 单输入方式改进为多传感器振动信号的多通道输入, 然后经过卷积、激活函数和最大池化对不同通道的传感器数据进行特征提取, 接着通过多通道特征融合模块对不同通道的特征信息进行加权融合, 最后使用卷积、池化操作提取深层特征信息. 此外在输出层之前, 使用了带残差连接的自注意力模块, 该模块能够关注网络中每一层的全局信息, 增强了对不同传感器振动信号的深层特征学习能力, 有效地提升网络故障识别准确率.

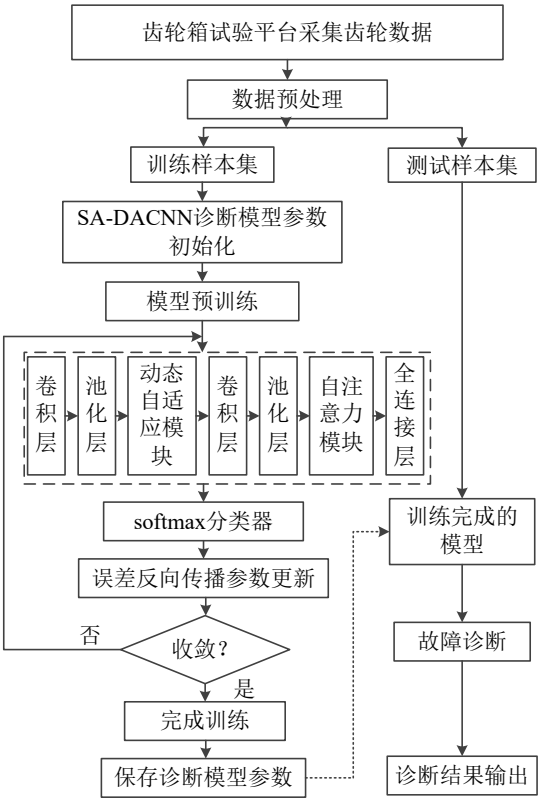


图 4 SA-DACNN 诊断逻辑流程图

图 4 为 SA-DACNN 的流程图, 其诊断步骤为:

(1) 采集齿轮箱不同位置的传感器信号;

(2) 对传感器信号进行数据处理, 划分数据为训练样本集、验证样本集和测试样本集, 其中训练样本集和验证样本集用于训练模型, 测试样本集用于验证模型;

(3) 模型参数初始化, 并将训练样本集和验证样本集输入模型, 通过误差反向传播算法优化模型参数;

(4) 判断模型是否收敛, 若收敛则完成训练并保存模型参数, 若不收敛则返回执行步骤 (3);

(5) 测试样本集输入已经训练完成的模型, 然后通过模型进行特征提取, 最终输出故障诊断结果.

3 实验仿真与结果分析

在本章节中, 使用两个齿轮箱数据评估所提方法的故障诊断效果. 实验的电脑配置为 AMD 锐龙 5-4600H 处理器, 实验框架为 keras-python.

3.1 案例 1

3.1.1 数据集描述

实验数据由东南大学动力传动系统动态模拟实验平台提供[23]. 该平台模拟了负载设置在 20Hz-0V 和 30Hz-2V 工况下, 5 种齿轮类型的实验数据. 5 种齿轮类型分别为齿轮断齿、齿轮缺口、齿面磨损、齿根裂纹和正常状态. 此外, 该平台采集了电机振动和扭矩、行星齿轮箱三个方向 (X、Y、Z) 和平行齿轮箱三个方向 (X、Y、Z) 共 8 个通道的实验数据. 在本案例中, 选取 4 个通道的实验数据, 每种齿轮故障类型取 600 样本, 并按照 3:1 比例划分训练集和测试集, 具体数据选取和划分表 1 所示.

表 1 东南大学数据集划分

工况	类别	训练集	测试集	标签
	齿轮断齿	450	150	0
	齿轮缺口	450	150	1
20Hz-0V	齿面磨损	450	150	2
	齿根裂纹	450	150	3
	正常状态	450	150	4
	齿轮断齿	450	150	5
	齿轮缺口	450	150	6
20Hz-0V	齿面磨损	450	150	7
	齿根裂纹	450	150	8
	正常状态	450	150	9

3.1.2 SA-DACNN 网络参数设计

SA-DACNN 网络在训练中采用 Adam 自适应优化算法, 在全连接层使用 Dropout 防止网络出现过拟合问题, 批次大小设置为 96, 循环迭代次数为 100, 模型的其它超参数设计见表 2 所示. 表 2 中 [batch, 2048, 1]×4 表示 4 个通道输入层的结构参数, 其中, 每个通道具有相同的结构参数, 2048 表示每个样本的采样点长度, 1 表示卷积操作的通道数.

表 2 网络结构参数

结构名称	结构参数	输出形状	参数量
输入层	[batch, 2048, 1]×4	[batch, 2048, 1]×4	0
卷积层 1	[32, 1, 16]×4	[batch, 128, 16]×4	528×4
激活层 1	Relu 激活函数	[batch, 128, 16]×4	0
池化层 1	[2, 1, 1]×4	[batch, 64, 16]×4	0
多通道特征融合模块	-	[batch, 64, 16]	0
卷积层 2	[3, 1, 32]	[batch, 64, 32]	1568
激活层 2	Relu 激活函数	[batch, 64, 32]	0
池化层 2	[2, 1, 1]	[batch, 32, 32]	0
卷积层 3	[3, 1, 64]	[batch, 32, 64]	6208
激活层 3	Relu 激活函数	[batch, 32, 64]	0
池化层 3	[2, 1, 1]	[batch, 16, 64]	0
自注意力层	[12]	[batch, 16, 12]	2304
全连接层	-	[batch, 100]	19300
Softmax 层	-	[batch, 10]	1010

3.1.3 可视化分析

为了更好地理解 SA-DACNN 特征提取过程, 本文把 4 个通道的实验数据输入到 SA-DACNN 中, 通过 t-SNE 技术将各层提取的特征降维可视化, 结果如图 5 所示, 由图 5 可知, 10 种原始数据刚开始输入网络时, 分布散乱无序、难以判别; 经过多通道特征融合模块后, 多数样本可以区分, 然而仍然存在一些故障样本误分的情况; 最后经过池化层、自注意力模块、全连接层和 Softmax 层处理后, 10 种故障样本被完全分离, 不同类型的故障样本不重叠.

3.1.4 多通道数据融合分析

为了分析 SA-DACNN 中多通道特征融合模块对多通道特征信息融合的效果, 本文进行单通道、2 通道和 3 通道输入实验, 具体实验设计和实验结果如表 3 所示, 由表 3 可知, 当输入为单通道时, 齿轮箱 X、Y 和 Z 轴的振动信号的诊断精确率高于电机端振动信号的诊断精确率, 这是因为电机端振动信号的幅值变化较大, 导致故障信息难以提取; 当输入为 2 通道时, 齿轮箱端的振动信号和电机端振动信号无论以何种组合, 其诊断精确率均高于单通道输入的诊断精确率, 其中, 当输入为齿轮箱 X 轴和齿轮箱 Y 轴时, 故障诊断精确率为 95.60%; 当输入为 3 通道时, 齿轮箱 X、Y 和 Z 轴的振动信号的诊断精确率已高达 99.20%. 通过以上分析可知, 随着输入通道数增加, 故障诊断精确率也逐渐增加, 这表明本文

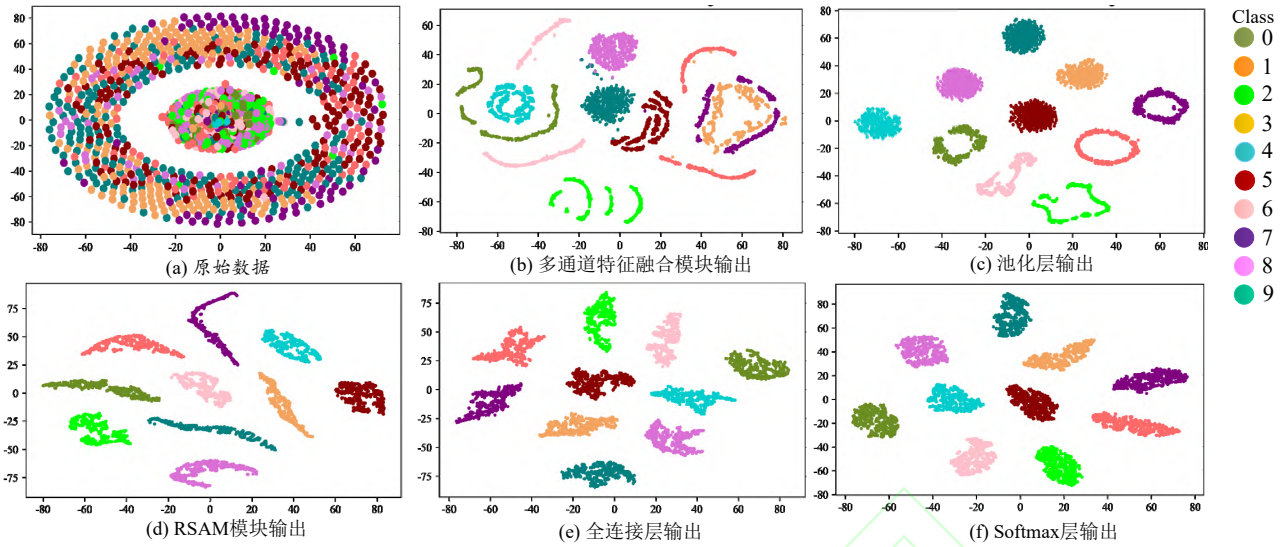


图 5　t-SNE 对各层特征降维可视化

设计的多通道特征融合模块能够有效地加权不同通道的权重, 突出重要通道的权重, 抑制其它干扰信息.

表 3 多通道数据融合诊断多指标结果

通道选择	信号类型	精确率	召回率	F1 均值
单通道	电机端信号	83.14%	84.34%	81.63%
	齿轮箱 X 轴	91.73%	91.36%	90.82%
	齿轮箱 Y 轴	91.40%	91.29%	90.77%
	齿轮箱 Z 轴	88.80%	90.27%	88.71%
2 通道	电机端信号、齿轮箱 X 轴	92.80%	93.68%	92.55%
	电机端信号、齿轮箱 Y 轴	92.77%	93.51%	92.37%
	电机端信号、齿轮箱 Z 轴	91.59%	91.89%	91.58%
	齿轮箱 X 轴、齿轮箱 Y 轴	95.60%	95.90%	95.60%
	齿轮箱 X 轴、齿轮箱 Z 轴	94.81%	94.59%	94.57%
	齿轮箱 Y 轴、齿轮箱 Z 轴	94.00%	94.59%	94.06%
3 通道	电机端信号、齿轮箱 X、Y 轴	98.33%	98.27%	98.32%
	电机端信号、齿轮箱 X、Z 轴	98.31%	98.19%	98.20%
	齿轮箱 X 轴、齿轮箱 Y、Z 轴	99.20%	99.17%	98.99%

3.1.5　不同方法性能比较

为了进一步验证 SA-DACNN 的故障诊断能力, 本文选取 2D-CNN[9]、HMS-MACNN[12]、MLPC-CNN[13]、P-2DCNN[24]、BP、LSTM、SA-CNN 和 DACNN 与本文所提方法进行比较, 对比方法具体描述为:(1)MLPC-CNN 为多层特征融合卷积神经网络, 该网络使用单传感器单通道卷积 (SSTSC), 处理所有的传感器数据, 使用两个卷积层, 在旁路分支使用平均池化, 学习率优化方法为 SGDM；(2)2D-CNN 网络由 2 个卷积层、2 个最大池化层和 2 个 FC 全连接层组成；(3)HMS-MACNN 为混合多尺度卷积神经网络, 该网络采用多尺度技术和自适应加权的多注意模块进行特征提取和故障识别, 网络由 2 个混合多尺度模块、3 个卷积层和 2 个注意力模块组成；(4)P-2DCNN 网络将不同位置传感器时域振动信号构造成二维矩阵, 网络由 2 个卷积层、2 个池化层组成；(5)BP 和 LSTM 选取是为了与 CNN 网络做对比；(6)SA-CNN 网络的结构参数与本文所提方法一样, 不同之处是网络中没有多通道特征融合模块；(7)DACNN 网络无自注意力模块, 其结构参数配置与本文方法一样.

表 4 不同方法的故障诊断结果

不同方法	训练准确率		测试准确率		训练速度
	平均值	标准差	平均值	标准差	(s/batch)
BP	95.49%	1.62%	74.93%	4.14%	10.95s
LSTM	96.90%	0.96%	84.33%	3.77%	45.32s
MLPC-CNN	99.27%	0.37%	97.47%	6.01%	4.95s
2D-CNN	99.81%	0.10%	95.96%	3.91%	6.51s
HMS-MACNN	99.52%	0.28%	98.17%	3.93%	19.89s
P-2DCNN	97.39%	0.86%	96.05%	3.46%	3.90s
SA-CNN	99.95%	0.12%	98.10%	4.23%	1.61s
DACNN	99.96%	0.03%	99.41%	1.34%	1.53s
SA-DACNN	100.0%	0.00%	100.0%	0.00%	1.70s

为了消除随机因素引起的干扰, 以上实验重复 10 次, 并采用 10 次实验的诊断准确率的平均值和标准差来评价各种方法的性能, 实验结果如表 4 所示. 由 4 表可知, 本文所提方法的训练准确率和测试准确率都高于对比方法, 本文所提方法的训练准确率和测试准确率为 100%, 并且每批次的训练时间为 1.7s, 这表明所提方法在故障诊断准确率和时效性方面具有一

定的优势. 相比于 BP 和 LSTM 的方法, 本文所提方法的测试准确率分别提高了 25.07% 和 15.67%, 表明卷积模型在特征提取和分类方面拥有其它网络无法比拟的优势. 相比于 MLPC-CNN 和 HMS-MACNN 的方法, 本文所提方法的测试集故障诊断精度分别提高了 2.53% 和 1.83%, 表明所设计的多通道特征融合模块能够对不同通道的传感器数据加权, 提升故障识别准确率. 相比于 2D-CNN 和 P-2DCNN 的方法, 本文所提方法的故障诊断精度分别提高了 4.04% 和 3.95%, 表明 1D 卷积网络更加适合振动信号的特征提取. 相比于 SA-CNN 和 DACNN, 本文所提方法的故障诊断精度分别提高了 1.90% 和 0.59%, 表明自注意力模块和多通道特征融合模块能够增强网络的故障辨识能力和泛化能力.

3.2 案例 2

3.2.1 数据集描述

实验数据来源于滚轴齿轮故障模拟试验平台. 该平台由三相交流异步电机、加载装置、测试轴承箱、传动机构、齿轮箱和变频控制器组成. 本案例被测对象为锥齿轮, 锥齿轮的故障状态分别为全齿断裂、小端断半齿、大端断半齿、均匀磨损和正常状态共 5 种不同的锥齿轮状态. 实验数据由安装在 9 点方向和 12 点方向上的剪切加速度传感器采集, 信号采用频率为 25.6KHz. 本文选取电机运行频率为 20Hz, 锥齿轮加载为 1HP 的实验数据, 每种锥齿故障样本数为 600, 按照 3:1 比例划分训练集和测试集, 数据集具体划分如表 5 所示.

表 5 滚轴齿轮数据集划分

类别	训练集	测试集	标签
齿轮大端断半齿	450	150	0
齿轮均匀磨损	450	150	1
齿轮全齿断裂	450	150	2
齿轮小端断半齿	450	150	3
齿轮正常	450	150	4

3.2.2 可视化分析

为了验证所提方法特征提取能力, 选用 t-SNE 技术将各个模块的特征进行可视化, 如图 6 所示. 由图 6 可以看到, 原始数据随机分布在二维可视化图中, 各个故障样本不能区分; 多通道特征融合模块能够让特征彼此分离, 相同的特征呈现聚集状态; Softmax 层将相同特征完全聚类, 不同特征完全分离. 通过可视化分析, 表明本文所提方法具有较好的故障识别能力和良好的聚类性能.

3.2.3 多通道数据融合分析

同案例 1 一样, 对案例 2 进行单通道和 2 通道输入实验, 实验设计和实验结果如表 6 所示, 由表 6 可知, 当输入为单通道时, 齿轮箱 X 轴和 Y 轴的诊断精确率分别为 97.86% 和 95.46%; 当输入为 2 通道时, 所提方法的诊断精确率已高达 99.73%, 这表明所设计的多通道特征融合模块能够对不同通道的传感器数据加权, 可以提升故障识别准确率. 此外, 为了更直观的显示诊断效果, 采用混淆矩阵对锥齿轮故障识别能力进一步分析, 结果如图 7 所示, 由图 7 可知, 单通道齿轮箱 X 轴输入时, 即 SA-DACNN1, 标签 2 和标签 3 对应故障类型为齿轮全齿断裂和齿轮小端断半齿, 存在 5% 的样本被误诊为其他类型故障. 相比之下, 本文所提方法除了将 1% 的齿轮小端断半齿被误诊为正常状态外, 其它样本的故障诊断准确率为 100%.

表 6 多通道数据融合诊断多指标结果

通道选择	信号类型	精确率	召回率	F1 均值
单通道	齿轮箱 X 轴	97.86%	97.97%	97.87%
	齿轮箱 Y 轴	95.46%	95.52%	95.37%
2 通道	齿轮箱 X 轴、齿轮箱 Y 轴	99.73%	99.72%	99.73%

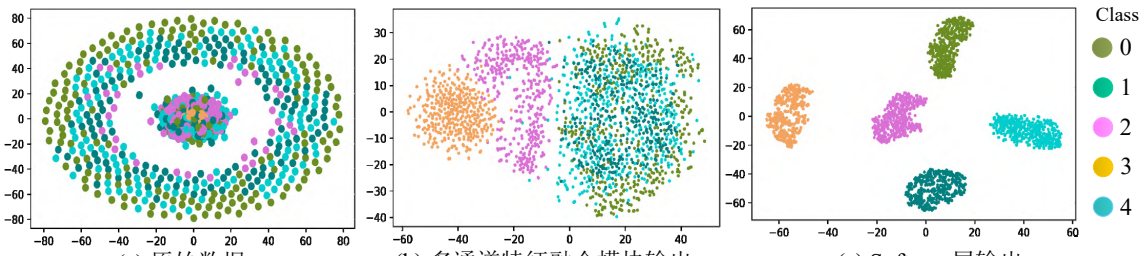
3.2.4 不同方法性能比较

与案例 1 相同, 本文仍然选用 BP、LSTM、MLPC-CNN、2D-CNN、HMS-MACNN、P-2DCNN、SA-CNN 和 DACNN 与所提方法 SA-DACNN 作对比, 对比结果如表 7 所示.

表 7 不同方法的故障诊断结果

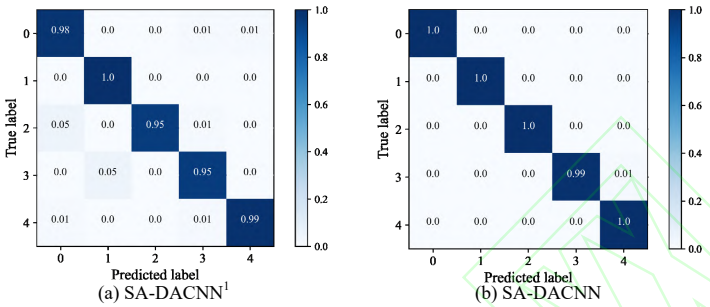
不同方法	训练准确率		测试准确率		训练速度
	平均值	标准差	平均值	标准差	(s/batch)
BP	96.81%	0.63%	74.53%	1.76%	5.07s
LSTM	97.42%	0.47%	84.42%	2.93%	23.2s
MLPC-CNN	99.90%	0.10%	98.00%	2.85%	1.43s
2D-CNN	99.76%	0.39%	97.86%	2.56%	2.47s
HMS-MACNN	99.32%	0.99%	96.69%	4.43%	3.43s
P-2DCNN	99.60%	0.38%	98.80%	0.62%	1.60s
SA-CNN	99.77%	0.36%	98.63%	2.57%	0.63s
DACNN	99.64%	0.45%	98.77%	2.31%	0.55s
SA-DACNN	99.94%	0.06%	99.64%	0.39%	0.71s

由表 7 可以看到, 所有方法的平均训练准确率均在 95% 以上, 其中 SA-DACNN 的训练准确率为 99.94%, 是所有方法中最好的; SA-DACNN 的平均测试准确率为 99.64%, 比 HMS-MACNN 方法高 2.95%, 比 2D-CNN 方法高 1.78%. 此外, 基于 CNN 模型的方法测试准确率明显高于 BP 和 LSTM, 这是因为 BP 和 LSTM 网络结构简单, 未采用先进的模块导致



(a) 原始数据 (b) 多通道特征融合模块输出 (c) Softmax层输出

图 6 t-SNE 对各层特征降维可视化



(a) SA-DACNN[1] (b) SA-DACNN

图 7 多通道数据融合混淆矩阵

特征提取能力不足; 与 SA-CNN 和 DACNN 方法相比, 去掉多通道特征融合模块后, 测试准确率下降了 1.01%, 去掉自注意力模块后, 测试准确率下降了 0.87%. 这表明本文所设计的多通道特征融合模块能够有效地加权不同通道的特征, 提高故障诊断精度; 自注意力模块的引入可以帮助网络提取到深层网络的重要特征. 通过以上分析可知, 本文所提方法的故障诊断准确率明显高于对比方法, 并且收敛速度更快.

4 结论

为了解决单一传感器数据下齿轮箱故障诊断效果差的问题, 本文提出了一种基于 SA-DACNN 的多传感器数据融合故障诊断方法, 并得到以下结论: (1) 本文设计的多通道特征融合模块可以融合不同通道的传感器数据, 并自适应地增强有用特征信息, 抑制干扰信息; 带残差连接的自注意力模块可以帮助网络更好地学习特征信息和标签之间的映射关系, 提升齿轮箱的故障诊断准确率. (2) 本文使用两个齿轮箱数据验证方法的有效性, 分析了单个传感器与多传感器数据对诊断效果的影响, 结果表明多传感器数据融合可以获得更加全面的故障特征, 有效提高故障识别精度. (3) 虽然本文方法可以解决多传感器数据融合问题, 但是齿轮箱通常运行在变工况环境下, 不同工况下故障诊断结果有所差异, 因此未来将针对变工况下齿轮箱多传感器数据融合做进一步研究.

参考文献 (References)

[1] Zhang L, Wang B, Liang P, et al. Semi-supervised fault diagnosis of gearbox based on feature pre-extraction mechanism and improved generative adversarial networks under limited labeled samples and noise environment[J]. Advanced Engineering Informatics, 2023, 58: 102211.

[2] Zhu Y, Pei Y, Wang A, et al. A partial domain adaptation scheme based on weighted adversarial nets with improved CBAM for fault diagnosis of wind turbine gearbox[J]. Engineering Applications of Artificial Intelligence, 2023, 125: 106674.

[3] 王进花, 岳亮辉, 曹洁等. 基于随机变分推理贝叶斯神经网络的发电机轴承故障诊断 [J]. 控制与决策,2023,38(04):1015-1021.
Wang Jinhua, Yue Lianghui, Cao Jie et al. Generator bearing fault diagnosis based on stochastic variational inference Bayesian neural network[J]. Control and Decision Making, 2023, 38(04): 1015-1021.

[4] 梁浩鹏, 曹洁, 赵小强. 基于 GADF 和 PAM-Resnet 的旋转机械小样本故障诊断方法 [J]. 控制与决策, 2023, 38(12): 3465-3472.
Liang Haopeng, Cao Jie, Zhao Xiaoqiang. A small-sample fault diagnosis method for rotating machinery based on GADF and PAM-Resnet[J]. Control and Decision Making, 2023, 38 (12): 3465-3472.

[5] Zhao X, Zhang Y. An intelligent diagnosis method of rolling bearing based on multi-scale residual shrinkage convolutional neural network[J]. Measurement Science and Technology, 2022, 33(8): 085103.

[6] Shao H, Lin J, Zhang L, et al. A novel approach of multisensory fusion to collaborative fault diagnosis in

maintenance[J]. Information Fusion, 2021, 74: 65-76.

[7] Liu K, Gebraeel N Z, Shi J. A data-level fusion model for developing composite health indices for degradation modeling and prognostic analysis[J]. IEEE Transactions on Automation Science and Engineering, 2013, 10(3): 652-664.

[8] Jing L, Wang T, Zhao M, et al. An adaptive multi-sensor data fusion method based on deep convolutional neural networks for fault diagnosis of planetary gearbox[J]. Sensors, 2017, 17(2): 414.

[9] Azamfar M, Singh J, Bravo-Imaz I, et al. Multisensor data fusion for gearbox fault diagnosis using 2-D convolutional neural network and motor current signature analysis[J]. Mechanical Systems and Signal Processing, 2020, 144: 106861.

[10] 杨洁, 万安平, 王景霖, 等. 基于多传感器融合卷积神经网络的航空发动机轴承故障诊断 [J]. 中国电机工程学报 2022, 42(13): 4933-4942.
Yang Jie,Wan Anping,Wang Jinglin et al. Aero-engine bearing fault diagnosis based on multi-sensor fusion convolutional neural network[J]. Chinese Journal of Electrical Engineering 2022, 42(13): 4933-4942.

[11] Gong W, Wang Y, Zhang M, et al. A fast anomaly diagnosis approach based on modified CNN and multisensor data fusion[J]. IEEE Transactions on Industrial Electronics, 2021, 69(12): 13636-13646.

[12] Xu Z, Bashir M, Zhang W, et al. An intelligent fault diagnosis for machine maintenance using weighted soft-voting rule based multi-attention module with multi-scale information fusion[J]. Information Fusion, 2022, 86: 17-29.

[13] Zhang Y, He L, Cheng G. MLPC-CNN: A multi-sensor vibration signal fault diagnosis method under less computing resources[J]. Measurement, 2022, 188: 110407.

[14] Jiao J, Zhao M, Lin J, et al. A comprehensive review on convolutional neural network in machine fault diagnosis[J]. Neurocomputing, 2020, 417: 36-63.

[15] 吴耀春, 赵荣珍, 靳伍银, 等. 面向数据不平衡的卷积神经网络故障辨识方法 [J]. 振动. 测试与诊断, 2022, 42(02): 299-307+408.
Wu Y.C., Zhao R.Z., Jin W.Y., et al. A data imbalance-oriented convolutional neural network fault identification method[J]. Vibration. Test and Diagnosis,2022,42(02):299-307+408.

[16] 姚齐水, 别帅帅, 余江鸿, 等. 一种结合改进 Inception V2 模块和 CBAM 的轴承故障诊断方法 [J]. 振动工程学报，2022，35(04):949-957.
Yao Qishui,Bu Shuaihua,Yu Jianghong et al. A bearing fault diagnosis method combining improved Inception V2 module and CBAM[J]. Journal of Vibration Engineering,2022,35(04):949-957.

[17] 宫文峰, 陈辉, 张泽辉, 等. 基于改进卷积神经网络的滚动轴承智能故障诊断研究 [J]. 振动工程学报, 2020, 33(02): 400-413.
Gong Wenfeng,Chen Hui,Zhang Zehui et al. Research on intelligent fault diagnosis of rolling bearings based on improved convolutional neural network[J]. Journal of Vibration Engineering,2020, 33(02): 400-413.

[18] 姚家琪, 荆华, 赵春晖. 一种面向噪声环境中旋转机械故障诊断的多模态耦合输入神经网络 [J]. 控制与决策, 2023, 38(07): 1918-1926.
Yao Jiaqi, Jing Hua, Zhao Chunhui. A multimodal coupled input neural network for fault diagnosis of rotating machinery in noisy environments[J]. Control and Decision Making, 2023, 38(07): 1918-1926.

[19] Wang Y, Huang J, Wang Y, et al. A CNN-based adaptive surface monitoring system for fused deposition modeling[J]. IEEE/ASME Transactions on Mechatronics, 2020, 25(5): 2287-2296.

[20] Zhang W, Li C, Peng G, et al. A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load[J]. Mechanical systems and signal processing, 2018, 100: 439-453.

[21] Xu D, Qiu H, Gao L, et al. A novel dual-stream self-attention neural network for remaining useful life estimation of mechanical systems[J]. Reliability Engineering & System Safety, 2022, 222: 108444.

[22] Liu S, Jiang H, Wu Z, et al. Data synthesis using deep feature enhanced generative adversarial networks for rolling bearing imbalanced fault diagnosis[J]. Mechanical Systems and Signal Processing, 2022, 163: 108139.

[23] Shao S, McAleer S, Yan R, et al. Highly accurate machine fault diagnosis using deep transfer learning[J]. IEEE Transactions on Industrial Informatics, 2018, 15(4): 2446-2455.

[24] Wang J, Wang D, Wang S, et al. Fault diagnosis of bearings based on multi-sensor information fusion and 2D convolutional neural network[J]. IEEE Access, 2021, 9: 23717-23725.

作者简介

张亚洲 (1996–), 男, 博士生, 从事旋转机械故障诊断与寿命预测等研究, E-mail: 1911599612@qq.com;

赵小强 (1969–), 男, 教授, 博士生导师, 从事故障诊断、图像处理、数据挖掘等研究, E-mail: xqzhao@lut.edu.cn;

惠永永(1992–), 男, 副教授, 学历, 从事间歇过程故障检测与诊断等研究, E-mail: 1048393569@qq.com;

陈鹏 (1992–), 男, 副教授, 学历, 从事旋转机械故障诊断等研究, E-mail: 1341987756@qq.com.

证书号第7427884号

专利公告信息

发明专利证书

发明名称：一种基于MSDC-Swin-T的齿轮箱故障诊断方法

专利权人：兰州理工大学

地　　址：730050 甘肃省兰州市七里河区兰工坪路287号

发 明 人：张亚洲;赵小强;徐蓉蓉;梁浩鹏;陈鹏;惠永永;柴靖轩
宋昭漾;牟淼;刘凯;强睿儒;徐锦涛

专 利 号：ZL 2023 1 1526466.0　　授权公告号：CN 117516925 B

专利申请日：2023年11月16日　　授权公告日：2024年10月08日

申请日时申请人：兰州理工大学

申请日时发明人：张亚洲;赵小强;徐蓉蓉;梁浩鹏;陈鹏;惠永永;柴靖轩
宋昭漾;牟淼;刘凯;强睿儒;徐锦涛

国家知识产权局依照中华人民共和国专利法进行审查，决定授予专利权，并予以公告。

专利权自授权公告之日起生效。专利权有效性及专利权人变更等法律信息以专利登记簿记载为准。

局长
申长雨

2024年10月08日